

ROBUST AND SCALABLE BAYES VIA A MEDIAN OF SUBSET POSTERIOR MEASURES

BY STANISLAV MINSKER^{*,†,‡}, SANVESH SRIVASTAVA^{*,‡,§}, LIZHEN
LIN^{*,‡} AND DAVID DUNSON^{*,‡}

Duke University[†] and SAMSI[§]

We propose a novel approach to Bayesian analysis that is provably robust to outliers in the data and often has computational advantages over standard methods. Our technique is based on splitting the data into non-overlapping subgroups, evaluating the posterior distribution given each independent subgroup, and then combining the resulting measures. The main novelty of our approach is the proposed aggregation step, which is based on the evaluation of a median in the space of probability measures equipped with a suitable collection of distances that can be quickly and efficiently evaluated in practice. We present both theoretical and numerical evidence illustrating the improvements achieved by our method.

1. Introduction. Contemporary data analysis problems pose several general challenges. One is resource limitations: massive data require computer clusters for storage and processing. Another problem occurs when data are severely contaminated by “outliers” that are not easily identified and removed. Following [Box and Tiao \(1968\)](#), an outlier can be defined as “*being an observation which is suspected to be partially or wholly irrelevant because it is not generated by the stochastic model assumed.*” While the topic of robust estimation has occupied an important place in the statistical literature for several decades and significant progress has been made in the theory of point estimation, robust Bayesian methods are not sufficiently well-understood.

Our main goal is to make a step towards solving these problems, proposing a general Bayesian approach that is (i) provably robust to the presence of

^{*}Authors were partially supported by grant R01-ES-017436 from the National Institute of Environmental Health Sciences (NIEHS) of the National Institutes of Health (NIH).

[†]Stanislav Minsker acknowledges support from NSF grants FODAVA CCF-0808847, DMS-0847388, ATD-1222567.

MSC 2010 subject classifications: Primary 62F15; secondary 68W15, 62G35

Keywords and phrases: Big data, Distributed computing, Geometric median, Parallel MCMC, Stochastic approximation, Wasserstein distance.

outliers in the data without any specific assumptions on their distribution or reliance on preprocessing; and (ii) scalable to big data sets through allowing computational algorithms to be implemented in parallel for different data subsets prior to an efficient aggregation step. The proposed method splits the sample into disjoint parts, implements Markov chain Monte Carlo (MCMC) or another posterior sampling method to obtain draws from each “subset posterior” in parallel, and then uses these draws to obtain weighted samples from the *median posterior (or M-posterior)*, a new probability measure which is a (properly defined) median of a collection of posterior distributions based on a partition of the data.

The paper is organized as follows: section 1.1 contains an overview of the existing literature and explains the goals that we aim to achieve in this work. Section 2 introduces the mathematical background and key facts used throughout the paper. Section 3 describes the main theoretical results for the median posterior. Section 4 presents details of algorithms, implementation, and numerical performance of the median posterior for several models. Proofs that are omitted in the main text are contained in the appendix.

1.1. *Discussion of related work.* A. Dasgupta remarks that (see the discussion following Berger (1994)): “Exactly what constitutes a study of Bayesian robustness is of course impossible to define.” The popular definition (which also indicates the main directions of research in this area) is due to J. Berger (Berger, 1994): “Robust Bayesian analysis is the study of the sensitivity of Bayesian answers to uncertain inputs. These uncertain inputs are typically the model, prior distribution, or utility function, or some combination thereof.” Outliers are typically accommodated by either employing heavy-tailed likelihoods (e.g., Svensen and Bishop (2005)) or by attempting to identify and remove them as a first step (as in Box and Tiao (1968) or Bayarri and Berger (1994)). The usual assumption in the Bayesian literature is that the distribution of the outliers can be modeled (e.g., using a t -distribution, contamination by a larger variance parametric distribution, etc). In this paper, we instead bypass the need to place a model on the outliers and do not require their removal prior to analysis. We also present evidence of robustness to model misspecification, see section 3.4.

Also relevant is recent progress in scalable Bayesian algorithms. Most methods designed for distributed computing share a common feature: they efficiently use the data subset available to a single machine and combine the “local” results for “global” learning, while minimizing communication among cluster machines (Smola and Narayanamurthy (2010)). A wide variety of

optimization-based approaches are available for distributed learning (Boyd et al., 2011); however, the number of similar Bayesian methods is limited. One of the reasons for this limitation is related to Markov chain Monte Carlo (MCMC), the dominating approach for approximating the posterior distribution of parameters in Bayesian models. While there are many efficient MCMC techniques for sampling from posterior distributions based on small subsets of the data (called “subset posteriors” in the sequel), to the best of our knowledge, there is no general rigorously justified approach for combining the subset posteriors into a single distribution for improved performance.

Three major approaches exist for scalable Bayesian learning in a distributed setting. The first approach independently evaluates the likelihood for each data subset across multiple machines and returns the likelihoods to a “master” machine, where they are appropriately combined with the prior using conditional independence assumptions of the probabilistic model. These two steps are repeated at every MCMC iteration (see Smola and Narayana-murthy (2010); Agarwal and Duchi (2012)). This approach is problem-specific and involves extensive communication among machines. The second approach uses a so-called stochastic approximation (SA) and successively learns “noisy” approximations to the full posterior distribution using data in small mini-batches. The accuracy of SA increases as it uses more data. A group of methods based on this approach uses sampling-based techniques to explore the posterior distribution through modified Hamiltonian or Langevin dynamics (e.g., Welling and Teh (2011); Ahn, Korattikara and Welling (2012); Korattikara, Chen and Welling (2013)). Unfortunately, these methods fail to accommodate discrete-valued parameters and multimodality. Another subgroup of methods uses deterministic variational approximations and learns the variational parameters of the approximated posterior through an optimization-based approach (see Wang, Paisley and Blei (2011); Hoffman et al. (2013); Broderick et al. (2013)). Although these techniques often have excellent predictive performance, it is well known (Bishop, 2006) that variational methods tend to substantially underestimate posterior uncertainty and provide a poor characterization of posterior dependence, while lacking theoretical guarantees.

Our approach instead falls in a third class of methods which avoid extensive communication among machines by running independent MCMC chains for each data subset and obtaining draws from subset posteriors. These subset posteriors can be combined in a variety of ways. Some of these methods simply average draws from each subset (Scott et al., 2013). Other alterna-

tives use an approximation to the full posterior distribution based on kernel density estimates (Neiswanger, Wang and Xing, 2013) or the so-called Weierstrass transform (Wang and Dunson, 2013). These methods have limitations related to the dimension of the parameter, moreover, their applicability and theoretical justification are restricted to parametric models. Unlike the method proposed below, none of the aforementioned algorithms are provably robust.

Our work was inspired by recent multivariate median-based techniques for robust estimation developed in Minsker (2013) (see also Hsu and Sabato (2013); Alon, Matias and Szegedy (1996); Lerasle and Oliveira (2011); Nemirovski and Yudin (1983) where similar ideas were applied in different frameworks).

2. Preliminaries. We proceed by recalling key definitions and facts which will be used throughout the paper.

2.1. Notation. In what follows, $\|\cdot\|_2$ denotes the standard Euclidean distance in \mathbb{R}^p and $\langle \cdot, \cdot \rangle_{\mathbb{R}^p}$ - the associated dot product. For a $p \times p$ matrix $A \in \mathbb{R}^{p \times p}$, $\text{Tr } A$ denotes its trace.

Given a totally bounded metric space (\mathbb{Y}, d) , the packing number $M(\varepsilon, \mathbb{Y}, d)$ is the maximal number N such that there exist N disjoint d -balls B_1, \dots, B_N of radius ε contained in \mathbb{Y} , i.e., $\bigcup_{j=1}^N B_j \subseteq \mathbb{Y}$. Let $\{p_\theta, \theta \in \Theta\}$ be a family of probability density functions on \mathbb{R}^p . Let $l, u : \mathbb{R}^p \mapsto \mathbb{R}$ be two functions such that $l(x) \leq u(x)$ for every $x \in \mathbb{R}^p$ and $d^2(l, u) := \int_{\mathbb{R}^p} (\sqrt{u} - \sqrt{l})^2(x) dx < \infty$.

A bracket $[l, u]$ consists of all functions $g : \mathbb{R}^p \mapsto \mathbb{R}$ such that $l(x) \leq g(x) \leq u(x)$ for all $x \in \mathbb{R}^p$. The bracketing number $N_{[]}(\varepsilon, \Theta, d)$ is the smallest number N such that there exist N brackets $[l_i, u_i]$, $i = 1, \dots, N$ satisfying $\{p_\theta, \theta \in \Theta\} \subseteq \bigcup_{i=1}^N [l_i, u_i]$ and $d(l_i, u_i) \leq \varepsilon$ for all $1 \leq i \leq N$.

For $y \in \mathbb{Y}$, δ_y denotes the Dirac measure concentrated at y . In other words, for any Borel-measurable B , $\delta_y(B) = I\{y \in B\}$, where $I\{\cdot\}$ is the indicator function. We will say that $k : \mathbb{Y} \times \mathbb{Y} \mapsto \mathbb{R}$ is a *kernel* if it is a symmetric, positive definite function. Assume that $(\mathbb{H}, \langle \cdot, \cdot \rangle_{\mathbb{H}})$ is a reproducing kernel Hilbert space (RKHS) of functions $f : \mathbb{Y} \mapsto \mathbb{R}$. Then k is a *reproducing kernel* for \mathbb{H} if for any $f \in \mathbb{H}$ and $y \in \mathbb{Y}$, $\langle f, k(\cdot, y) \rangle_{\mathbb{H}} = f(y)$ (see Aronszajn (1950) for details).

For a square-integrable function $f \in L_2(\mathbb{R}^p)$, \hat{f} stands for its Fourier transform. For $x \in \mathbb{R}$, $\lfloor x \rfloor$ denotes the largest integer not greater than x . Other objects and definitions are introduced in the course of exposition when necessity arises.

2.2. Generalizations of the univariate median. Let \mathbb{Y} be a normed space with norm $\|\cdot\|$, and let μ be a probability measure on $(\mathbb{Y}, \|\cdot\|)$ equipped with Borel σ -algebra. Define the *geometric median* of μ by

$$x_* = \operatorname{argmin}_{y \in \mathbb{Y}} \int_{\mathbb{Y}} (\|y - x\| - \|x\|) \mu(dx).$$

In this paper, we focus on the special case when μ is a uniform distribution on a finite collection of atoms $x_1, \dots, x_m \in \mathbb{Y}$, so that

$$(2.1) \quad x_* = \operatorname{med}_g(x_1, \dots, x_m) := \operatorname{argmin}_{y \in \mathbb{Y}} \sum_{j=1}^m \|y - x_j\|.$$

Geometric median exists under rather general conditions; for example, if \mathbb{Y} is a Hilbert space (this case will be our main focus, for more general conditions see [Kempman \(1987\)](#)). Moreover, it is well-known that in this situation $x_* \in \operatorname{co}(x_1, \dots, x_m)$ – the convex hull of x_1, \dots, x_m (meaning that there exist nonnegative α_j , $j = 1 \dots m$, $\sum_{j=1}^m \alpha_j = 1$ such that $x_* = \sum_{j=1}^m \alpha_j x_j$).

Another useful generalization of the univariate median is defined as follows. Let (\mathbb{Y}, d) be a metric space with metric d , and $x_1, \dots, x_m \in \mathbb{Y}$. Define B_* to be the d -ball of minimal radius such that it is centered at one of $\{x_1, \dots, x_m\}$ and contains at least half of these points. Then the median $\operatorname{med}_0(x_1, \dots, x_m)$ of x_1, \dots, x_m is the center of B_* . In other words, let

$$(2.2) \quad \varepsilon_* := \inf \left\{ \varepsilon > 0 : \exists j = j(\varepsilon) \in \{1, \dots, m\} \text{ and } I(j) \subset \{1, \dots, m\} \text{ such that } |I(j)| > \frac{m}{2} \text{ and } \forall i \in I(j), d(x_i, x_j) \leq 2\varepsilon \right\},$$

$j_* := j(\varepsilon_*)$, where ties are broken arbitrarily, and set

$$(2.3) \quad x_* = \operatorname{med}_0(x_1, \dots, x_m) := x_{j_*}.$$

We will say that x_* is the *metric median* of x_1, \dots, x_m . Note that x_* always belongs to $\{x_1, \dots, x_m\}$. Advantages of this definition are its generality (only metric space structure is assumed) and simplicity of numerical

evaluation since only the pairwise distances $d(x_i, x_j)$, $i, j = 1, \dots, m$ are required to compute the median. This construction was previously employed in [Nemirovski and Yudin \(1983\)](#) in the context of stochastic optimization. A closely related notion of the median was used in [Lopuhaa and Rousseeuw \(1991\)](#) under the name of the “minimal volume ellipsoid” estimator.

Finally, we recall an important property of the median (shared both by med_g and med_0) which states that it transforms a collection of independent, “weakly concentrated” estimators into a single estimator with significantly stronger concentration properties. Given q, α such that $0 < q < \alpha < 1/2$, define

$$(2.4) \quad \psi(\alpha, q) := (1 - \alpha) \log \frac{1 - \alpha}{1 - q} + \alpha \log \frac{\alpha}{q}.$$

The following result is a version of Theorem 3.1 in [Minsker \(2013\)](#):

THEOREM 2.1.

a Assume that $(\mathbb{H}, \|\cdot\|)$ is a Hilbert space and $\theta_0 \in \mathbb{H}$. Let $\hat{\theta}_1, \dots, \hat{\theta}_m \in \mathbb{H}$ be a collection of independent random variables. Let the constants α, q, γ be such that $0 < q < \alpha < 1/2$, and $0 \leq \gamma < \frac{\alpha - q}{1 - q}$. Suppose $\varepsilon > 0$ is such that for all j , $1 \leq j \leq \lfloor (1 - \gamma)m \rfloor + 1$,

$$(2.5) \quad \Pr \left(\|\hat{\theta}_j - \theta_0\| > \varepsilon \right) \leq q.$$

Let $\hat{\theta}_* = \text{med}_g(\hat{\theta}_1, \dots, \hat{\theta}_m)$ be the geometric median of $\{\hat{\theta}_1, \dots, \hat{\theta}_m\}$. Then

$$\Pr \left(\|\hat{\theta}_* - \theta_0\| > C_\alpha \varepsilon \right) \leq e^{-m(1-\gamma)\psi\left(\frac{\alpha-\gamma}{1-\gamma}, q\right)},$$

where $C_\alpha = (1 - \alpha)\sqrt{\frac{1}{1-2\alpha}}$.

b Assume that (\mathbb{Y}, d) is a metric space and $\theta_0 \in \mathbb{Y}$. Let $\hat{\theta}_1, \dots, \hat{\theta}_m \in \mathbb{Y}$ be a collection of independent random variables. Let the constants q, γ be such that $0 < q < \frac{1}{2}$ and $0 \leq \gamma < \frac{1/2-q}{1-q}$. Suppose $\varepsilon > 0$ are such that for all j , $1 \leq j \leq \lfloor (1 - \gamma)m \rfloor + 1$,

$$(2.6) \quad \Pr \left(d(\hat{\theta}_j, \theta_0) > \varepsilon \right) \leq q.$$

Let $\hat{\theta}_* = \text{med}_0(\hat{\theta}_1, \dots, \hat{\theta}_m)$. Then

$$\Pr \left(d(\hat{\theta}_*, \theta_0) > 3\varepsilon \right) \leq e^{-m(1-\gamma)\psi\left(\frac{1/2-\gamma}{1-\gamma}, q\right)}.$$

PROOF. See Appendix A. \square

Theorem 2.1 implies that the concentration of the geometric median of independent estimators around the true parameter value improves geometrically fast with respect to the number of such estimators, while the estimation rate is preserved. In our case, these estimators will be the posterior distributions based on disjoint subsets of observations. Parameter γ allows to take the corrupted observations into account: if the initial sample contains not more than $\lfloor \gamma m \rfloor$ outliers (of arbitrary nature), then at most $\lfloor \gamma m \rfloor$ estimators amongst $\{\theta_1, \dots, \theta_m\}$ can be affected.

2.3. Distances between probability measures. Next, we discuss the special family of distances between probability measures that will be used throughout the paper. These distances provide the necessary structure to evaluate the medians in the space of measures, as discussed above. Since our goal is to develop computationally efficient techniques, we only consider distances that admit accurate numerical approximation.

Assume that (\mathbb{X}, ρ) is a separable metric space, and let $\mathcal{F} = \{f : \mathbb{X} \mapsto \mathbb{R}\}$ be a collection of real-valued functions. Given two Borel probability measures P, Q on \mathbb{X} , define

$$(2.7) \quad \|P - Q\|_{\mathcal{F}} := \sup_{f \in \mathcal{F}} \left| \int_{\mathbb{X}} f(x) d(P - Q)(x) \right|.$$

Important special cases include the situation when

$$(2.8) \quad \mathcal{F} = \mathcal{F}_L := \{f : \Theta \mapsto \mathbb{R} \text{ s.t. } \|f\|_L \leq 1\},$$

where $\|f\|_L := \sup_{x_1 \neq x_2} \frac{|f(x_1) - f(x_2)|}{\rho(x_1, x_2)}$ is the Lipschitz constant of f .

It is well-known (Dudley (2002), Theorem 11.8.2) that in this case $\|P - Q\|_{\mathcal{F}_L}$ is equal to the Wasserstein distance (also known as the Kantorovich-Rubinstein distance)

$$(2.9) \quad d_{W_{1,\rho}}(P, Q) = \inf \left\{ \mathbb{E} \rho(\mathbf{X}, \mathbf{Y}) : \mathcal{L}(\mathbf{X}) = P, \mathcal{L}(\mathbf{Y}) = Q \right\},$$

where $\mathcal{L}(\mathbf{Z})$ denotes the law of a random variable \mathbf{Z} and the infimum on the right is taken over the set of all joint distributions of (\mathbf{X}, \mathbf{Y}) with marginals P and Q . When the underlying metric ρ is clear from the context, we will simply write $d_{W_1}(P, Q)$ in what follows.

Another fruitful structure emerges when \mathcal{F} is a unit ball in a Reproducing Kernel Hilbert Space $(\mathbb{H}, \langle \cdot, \cdot \rangle_{\mathbb{H}})$ with a reproducing kernel $k : \mathbb{X} \times \mathbb{X} \mapsto \mathbb{R}$. That is,

$$(2.10) \quad \mathcal{F} = \mathcal{F}_k := \{f : \mathbb{X} \mapsto \mathbb{R}, \|f\|_{\mathbb{H}} := \sqrt{\langle f, f \rangle_{\mathbb{H}}} \leq 1\}.$$

Let $\mathcal{P}_k := \{P \text{ is a probability measure, } \int_{\mathbb{X}} \sqrt{k(x, x)} dP(x) < \infty\}$, and assume that $P, Q \in \mathcal{P}_k$. Theorem 1 in [Sriperumbudur et al. \(2010\)](#) implies that the corresponding distance between measures P and Q takes the form

$$(2.11) \quad \|P - Q\|_{\mathcal{F}_k} = \left\| \int_{\mathbb{X}} k(x, \cdot) d(P - Q)(x) \right\|_{\mathbb{H}}.$$

It follows that $P \mapsto \int_{\mathbb{X}} k(x, \cdot) dP(x)$ is an embedding of \mathcal{P}_k into the Hilbert space \mathbb{H} which can be seen as an application of the “kernel trick” in our setting. The Hilbert space structure allows one to use fast numerical methods to approximate the geometric median, see section 4 below.

REMARK 2.2. Note that when P and Q are discrete measures (e.g., $P = \sum_{j=1}^{N_1} \beta_j \delta_{z_j}$ and $Q = \sum_{j=1}^{N_2} \gamma_j \delta_{y_j}$), then

$$(2.12) \quad \|P - Q\|_{\mathcal{F}_k}^2 = \sum_{i,j=1}^{N_1} \beta_i \beta_j k(z_i, z_j) + \sum_{i,j=1}^{N_2} \gamma_i \gamma_j k(y_i, y_j) - 2 \sum_{i=1}^{N_1} \sum_{j=1}^{N_2} \beta_i \gamma_j k(z_i, y_j).$$

In this paper, we will only consider *characteristic* kernels, which means that $\|P - Q\|_{\mathcal{F}_k} = 0$ if and only if $P = Q$. It follows from Theorem 7 in [Sriperumbudur et al. \(2010\)](#) that a sufficient condition for k to be characteristic is its *strict positive definiteness*: we say that k is *strictly positive definite* if it is bounded, measurable, and such that for all non-zero signed Borel measures ν

$$\iint_{\mathbb{X} \times \mathbb{X}} k(x, y) d\nu(x) d\nu(y) > 0.$$

When $\mathbb{X} = \mathbb{R}^p$, a simple sufficient criterion for the kernel k to be characteristic follows from Theorem 9 in [Sriperumbudur et al. \(2010\)](#):

PROPOSITION 2.3. *Let $\mathbb{X} = \mathbb{R}^p$, $p \geq 1$. Assume that $k(x, y) = \phi(x - y)$ for some bounded, continuous, integrable, positive-definite function $\phi : \mathbb{R}^p \mapsto \mathbb{R}$.*

1. Let $\widehat{\phi}$ be the Fourier transform of ϕ . If $|\widehat{\phi}(x)| > 0$ for all $x \in \mathbb{R}^p$, then k is characteristic;
2. If ϕ is compactly supported, then k is characteristic.

REMARK 2.4. It is important to mention that in practical applications, we (almost) always deal with *empirical measures* based on a collection of samples from the posterior distributions. A natural question is the following: if P and Q are probability measures on \mathbb{R}^D and P_n, Q_m are their empirical versions, what is the size of the error

$$e_{m,n} := \left| \|P - Q\|_{\mathcal{F}_k} - \|P_m - Q_n\|_{\mathcal{F}_k} \right|?$$

A useful and favorable fact is that $e_{m,n}$ often does not depend on D : under weak assumptions on kernel k , $e_{n,m}$ has an upper bound of order $m^{-1/2} + n^{-1/2}$ (see Corollary 12 in [Sriperumbudur et al. \(2009\)](#)). On the other hand, the bound for the (stronger) Wasserstein distance is not dimension-free and is of order $m^{-1/(D+1)} + n^{-1/(D+1)}$.

Finally, we recall the definition of the well-known Hellinger distance. Assume that P and Q are probability measures on \mathbb{R}^D which are absolutely continuous with respect to Lebesgue measure with densities p and q respectively. Then the Hellinger distance between P and Q is given by

$$h(P, Q) := \sqrt{\frac{1}{2} \int_{\mathbb{R}^D} \left(\sqrt{p(x)} - \sqrt{q(x)} \right)^2 dx}.$$

3. Contributions and main results. This section explains the construction of “median posterior” (or M-posterior) distribution, along with the theoretical guarantees for its performance.

3.1. *Construction of robust posterior distribution.* Let $\{P_\theta, \theta \in \Theta\}$ be a family of probability distributions over \mathbb{R}^D indexed by Θ . Suppose that for all $\theta \in \Theta$, P_θ is absolutely continuous with respect to Lebesgue measure dx on \mathbb{R}^D with $dP_\theta(\cdot) = p_\theta(\cdot)dx$. In what follows, we equip Θ with a “Hellinger metric”

$$(3.1) \quad \rho(\theta_1, \theta_2) := h(P_{\theta_1}, P_{\theta_2}),$$

and assume that the metric space (Θ, ρ) is separable.

Let X_1, \dots, X_n be i.i.d. \mathbb{R}^D -valued random vectors defined on a probability space (Ω, \mathcal{B}, P) with unknown distribution $P_0 := P_{\theta_0}$ for some $\theta_0 \in \Theta$. Bayesian inference of P_0 requires specifying a *prior* distribution Π over Θ (equipped with the Borel σ -algebra induced by ρ). The *posterior* distribution given the observations $\mathcal{X}_n := \{X_1, \dots, X_n\}$ is a random probability measure on Θ defined by

$$\Pi_n(B|\mathcal{X}_n) := \frac{\int_B \prod_{i=1}^n p_\theta(X_i) d\Pi(\theta)}{\int_\Theta \prod_{i=1}^n p_\theta(X_i) d\Pi(\theta)}$$

for all Borel measurable sets $B \subseteq \Theta$. It is known (see [Ghosal, Ghosh and Van Der Vaart \(2000\)](#)) that under rather general assumptions the posterior distribution Π_n “contracts” towards θ_0 , meaning that

$$\Pi_n(\theta \in \Theta : \rho(\theta, \theta_0) \geq \varepsilon_n | \mathcal{X}_n) \rightarrow 0$$

almost surely or in probability as $n \rightarrow \infty$ for a suitable sequence $\varepsilon_n \rightarrow 0$.

One of the questions that we address can be formulated as follows: what happens if some observations in \mathcal{X}_n are corrupted, e.g., if \mathcal{X}_n contains outliers of arbitrary nature and magnitude? In this case, the usual posterior distribution might concentrate “far” from the true value θ_0 , depending on the amount of corruption in the sample.

We proceed with a general description of our proposed algorithm for constructing a robust version of the posterior distribution. Let $1 \leq m \leq n/2$ be an integer. Divide the sample \mathcal{X}_n into m disjoint groups G_1, \dots, G_m of size $|G_j| \geq \lfloor n/m \rfloor$ each:

$$\{X_1, \dots, X_n\} = \bigcup_{j=1}^m G_j, \quad G_i \cap G_l = \emptyset \text{ for } i \neq j, \quad |G_j| \geq \lfloor n/m \rfloor, \quad j = 1 \dots m.$$

A typical choice of m is $m \simeq \log n$, so that the groups G_j are sufficiently large (however, other choices are possible as well).

Let Π be a prior distribution over Θ , and let $\{\Pi_n^{(j)}(\cdot) := \Pi_n(\cdot | G_j), \quad j = 1, \dots, m\}$ be the family of subset posterior distributions depending on disjoint subgroups G_j , $j = 1, \dots, m$:

$$\Pi_n(B|G_j) := \frac{\int_B \prod_{i \in G_j} p_\theta(X_i) d\Pi(\theta)}{\int_\Theta \prod_{i \in G_j} p_\theta(X_i) d\Pi(\theta)}.$$

Define the median posterior (or M-posterior) as

$$(3.2) \quad \hat{\Pi}_{n,g} := \text{med}_g(\Pi_n^{(1)}, \dots, \Pi_n^{(m)}),$$

or

$$(3.3) \quad \hat{\Pi}_{n,0} := \text{med}_0(\Pi_n^{(1)}, \dots, \Pi_n^{(m)}),$$

where the medians $\text{med}_g(\cdot)$ and $\text{med}_0(\cdot)$ are evaluated with respect to $\|\cdot\|_{\mathcal{F}_L}$ or $\|\cdot\|_{\mathcal{F}_k}$ introduced in section 2.2 above. Note that $\hat{\Pi}_{n,g}$ and $\hat{\Pi}_{n,0}$ are always probability measures: indeed, due to the aforementioned properties of a geometric median, there exists $\alpha_1 \geq 0, \dots, \alpha_m \geq 0$, $\sum_{j=1}^m \alpha_j = 1$ such that $\hat{\Pi}_{n,g} = \sum_{j=1}^m \alpha_j \Pi_n^{(j)}$, and $\hat{\Pi}_{n,0} \in \{\Pi_n^{(1)}(\cdot), \dots, \Pi_n^{(m)}(\cdot)\}$ by definition.

REMARK 3.1.

1. In the situation when X_1, \dots, X_n are not identically distributed (e.g., for regression problem with fixed design on a uniform grid), our method can still be applied, but the partition scheme for groups G_j needs to be defined carefully. In particular, for regression with fixed design, each subgroup can be chosen to contain the data that still “lives” on a uniform grid, but with larger mesh size.
2. Note that, in principle, the prior distribution can be different for every subgroup, but we do not discuss this possibility here in detail.
3. In practical applications, “small” weights α_j in the representation $\hat{\Pi}_{n,g} = \sum_{j=1}^m \alpha_j \Pi_n^{(j)}$ are set to 0 (and the remaining weights are properly rescaled) for improved performance, see step 3 of Algorithm 2.

While $\hat{\Pi}_{n,g}$ and $\hat{\Pi}_{n,0}$ possess several nice properties (such as robustness to outliers), in practice they often overestimate the uncertainty about θ_0 , especially when the number of groups m is large. To overcome this difficulty, we suggest a modification of our approach where the random measures $\Pi_n^{(j)}$ are replaced by the *stochastic approximations* $\Pi_{n,m}(\cdot|G_j)$, $j = 1, \dots, m$ of the full posterior distribution. To this end, define the stochastic approximation based on the subsample G_j as

$$(3.4) \quad \Pi_{n,m}(B|G_j) := \frac{\int_B \left(\prod_{i \in G_j} p_\theta(X_i) \right)^m d\Pi(\theta)}{\int_\Theta \left(\prod_{i \in G_j} p_\theta(X_i) \right)^m d\Pi(\theta)}.$$

In other words, $\Pi_{n,m}(\cdot|G_j)$ is obtained as a posterior distribution given that each data point from G_j is observed m times. While each of $\Pi_{n,k}(\cdot|G_j)$ might be “unstable”, the geometric median $\hat{\Pi}_{n,g}^{\text{st}}$ (or the metric median $\hat{\Pi}_{n,0}^{\text{st}}$) of these random measures improves stability and yields smaller credible sets with good coverage properties. Practical performance of $\hat{\Pi}_{n,g}^{\text{st}}$ is often superior as compared to $\hat{\Pi}_{n,g}$ in our experiments. In all numerical simulations below, we use stochastic approximations and $\hat{\Pi}_{n,g}^{\text{st}}$ unless noted otherwise.

REMARK 3.2. The reader might question the necessity of studying both $\hat{\Pi}_{n,g}$ and $\hat{\Pi}_{n,0}$ simultaneously. While the geometric median showed better average performance in our numerical simulations as compared to metric median, its efficient evaluation requires the Hilbert space structure on the space of measures, while the “metric median” can be evaluated directly with respect to the Wasserstein distance. The latter might be preferable in some cases (e.g., in nonparametric models when construction of the kernel is difficult). In addition, the result does not depend on the choice of the kernel which can be appealing in some settings.

3.2. *Convergence of posterior distribution and applications to robust Bayesian inference.* Our first result establishes the “weak concentration” property of the posterior distribution around the true parameter. Let $\delta_0 := \delta_{\theta_0}$ be the Dirac measure supported on $\theta_0 \in \Theta$. The following statement is similar in spirit to Theorem 2.1 in Ghosal, Ghosh and Van Der Vaart (2000), the main difference being that we are interested in the Wasserstein distance $d_{W_{1,\rho}}(\Pi_n(\cdot|\mathcal{X}_l), \delta_0)$ rather than the contraction rate of posterior distribution. Here, the Wasserstein distance is evaluated with respect to the metric space structure of (Θ, ρ) , where ρ is the “Hellinger metric” defined in (3.1).

THEOREM 3.3. *Let $\mathcal{X}_l = \{X_1, \dots, X_l\}$ be an i.i.d. sample from P_0 . Assume that $\varepsilon_l > 0$ and $\Theta_l \subset \Theta$ are such that for a universal constant $K > 0$ and some constant $C > 0$*

- (1) *the packing number satisfies $\log M(\varepsilon_l, \Theta_l, \rho) \leq l\varepsilon_l^2$,*
- (2) *$\Pi(\Theta \setminus \Theta_l) \leq \exp(-l\varepsilon_l^2(C + 4))$,*
- (3) *$\Pi\left(\theta : -P_0\left(\log \frac{p_\theta}{p_0}\right) \leq \varepsilon_l^2, P_0\left(\log \frac{p_\theta}{p_0}\right)^2 \leq \varepsilon_l^2\right) \geq \exp(-Cl\varepsilon_l^2)$,*
- (4) *$e^{-\tilde{K}l\varepsilon_l^2} \leq \varepsilon_l$, where $\tilde{K} = \min(K/2, 1)$.*

Then there exists $R = R(C, K)$ such that

$$(3.5) \quad \Pr \left(d_{W_1}(\delta_0, \Pi_l(\cdot | \mathcal{X}_l)) \geq R\varepsilon_l \right) \leq \frac{1}{l\varepsilon_l^2} + 4e^{-(1+K/2)l\varepsilon_l^2/2}.$$

PROOF. The proof mimics the argument behind Theorem 2.1 in [Ghosal, Ghosh and Van Der Vaart \(2000\)](#), with several modifications. For reader's convenience, details are outlined in [Appendix B](#). \square

Combination of Theorems [3.3](#) and [2.1](#) yields the following inequality for $\hat{\Pi}_{n,0}$.

COROLLARY 3.4. *Let X_1, \dots, X_n be an i.i.d. sample from P_0 , and assume that $\hat{\Pi}_{n,0}$ is defined with respect to the Wasserstein distance $\|\cdot\|_{\mathcal{F}_L}$ as in [\(3.3\)](#) above. Set $l := \lfloor n/m \rfloor$, assume that conditions of Theorem [3.3](#) hold, and, moreover, that ε_l satisfies*

$$q := \frac{1}{l\varepsilon_l^2} + 4e^{-(1+K/2)l\varepsilon_l^2/2} < \frac{1}{2}.$$

Then

$$\Pr \left(d_{W_1}(\delta_0, \hat{\Pi}_{n,0}) \geq 3R\varepsilon_l \right) \leq e^{-m\psi(1/2,q)}.$$

PROOF. It is enough to apply part (b) of Theorem [2.1](#) with $\gamma = 0$ to the independent random measures $\Pi_n(\cdot | G_j)$, $j = 1, \dots, m$. Note that the “weak concentration” assumption [\(2.6\)](#) is implied by [\(3.5\)](#). \square

Once again, note the exponential improvement of concentration as compared to Theorem [3.3](#). At the same time, the rate ε_l is typically only slightly worse than the rate of contraction of the full posterior distribution: if $m \simeq \log n$, then the difference is only logarithmic.

REMARK 3.5. The case when the sample $\mathcal{X}_n = \{X_1, \dots, X_n\}$ contains $\lfloor \gamma m \rfloor$ outliers (which are arbitrary random variables not necessarily sampled from P_0) for some $\gamma < q$ can be handled similarly. This more general bound is readily implied by Theorem [2.1](#). In most examples throughout the paper, we consider the case $\gamma = 0$ for simplicity since the generalization is a trivial corollary of Theorem [2.1](#).

It is worth noticing that the conclusion of Theorem 3.3 can be strengthened. Namely, one can obtain exponential concentration bounds for the usual posterior distribution (the case $m = 1$) if condition (3) of Theorem 3.3 is replaced by a more restrictive bound (see Theorems 2.2 and 2.3 in Ghosal, Ghosh and Van Der Vaart (2000)). However, the question of robustness to outliers is not addressed in this situation.

While the result of the previous statement is promising, numerical approximation and sampling from the “robust posterior” $\hat{\Pi}_{n,0}$ is problematic due to the underlying geometry defined by the Hellinger metric and the associated Wasserstein distance that is often hard to estimate in practice. Our next goal is to derive similar guarantees for the geometric and metric medians evaluated with respect to the computationally tractable distance measure that is induced by embedding the corresponding posterior distributions in a Hilbert space as discussed in section 2.3 above.

Let k be a characteristic kernel defined on $\Theta \times \Theta$. Kernel k defines a metric on Θ

$$(3.6) \quad \rho_k(\theta_1, \theta_2) := \|k(\cdot, \theta_1) - k(\cdot, \theta_2)\|_{\mathbb{H}} = \left(k(\theta_1, \theta_1) + k(\theta_2, \theta_2) - 2k(\theta_1, \theta_2) \right)^{1/2},$$

where \mathbb{H} is the RKHS associated to k . We will assume that (Θ, ρ_k) is separable.

Let $f \in \mathbb{H}$ and note that, due to the reproducing property and Cauchy-Schwarz inequality, we have

$$(3.7) \quad \begin{aligned} f(\theta_1) - f(\theta_2) &= \langle f, k(\cdot, \theta_1) - k(\cdot, \theta_2) \rangle_{\mathbb{H}} \\ &\leq \|f\|_{\mathbb{H}} \|k(\cdot, \theta_1) - k(\cdot, \theta_2)\|_{\mathbb{H}} = \|f\|_{\mathbb{H}} \rho_k(\theta_1, \theta_2). \end{aligned}$$

Therefore, $\mathcal{F}_k \subseteq \mathcal{F}_L$ and $\|P - Q\|_{\mathcal{F}_k} \leq \|P - Q\|_{\mathcal{F}_L}$, where \mathcal{F}_k and \mathcal{F}_L were defined in (2.10) and (2.8) respectively, and the underlying metric structure is given by ρ_k . Hence, convergence with respect to $\|\cdot\|_{\mathcal{F}_L}$ implies convergence with respect to $\|\cdot\|_{\mathcal{F}_k}$.

As we have already observed, $\|P - Q\|_{\mathcal{F}_k}$ is equal to the Wasserstein distance $d_{W_{1, \rho_k}}(P, Q)$, where the underlying metric space (Θ, ρ_k) is equipped with the new metric ρ_k instead of the Hellinger metric. The following assumption allows us to translate results of Theorem 3.3 into the new setting.

ASSUMPTION 3.6. *Let $h(P_{\theta_1}, P_{\theta_2})$ be the Hellinger distance between P_{θ_1} and P_{θ_2} . Assume there exist positive constants γ and \tilde{C} such that for all*

$\theta_1, \theta_2 \in \Theta$,

$$h(P_{\theta_1}, P_{\theta_2}) \geq \tilde{C} \rho_k^\gamma(\theta_1, \theta_2).$$

EXAMPLE 3.7. It is well-known that the Hellinger distance between two multivariate normal distributions $P_1 = N(\mu_1, \Sigma_1)$ and $P_2 = N(\mu_2, \Sigma_2)$ is equal to

$$h^2(P_1, P_2) = 1 - \frac{\det^{1/4}(\Sigma_1 \Sigma_2)}{\det^{1/2}\left(\frac{\Sigma_1 + \Sigma_2}{2}\right)} \exp\left(-\frac{1}{8} \Delta\mu^T \left(\frac{\Sigma_1 + \Sigma_2}{2}\right)^{-1} \Delta\mu\right),$$

where $\Delta\mu = \mu_1 - \mu_2$. In particular, it implies that for the family $\{P_\theta = N(\theta, \Sigma), \theta \in \mathbb{R}^D\}$ with $\Sigma \succ 0$ and the kernel

$$k(\theta_1, \theta_2) := \exp\left(-\frac{1}{8}(\theta_1 - \theta_2)^T \Sigma^{-1}(\theta_1 - \theta_2)\right),$$

assumption 3.6 holds with $\tilde{C} = \frac{1}{\sqrt{2}}$ and $\gamma = 1$ (moreover, it holds with equality).

This can be extended to the case of general exponential families as follows. Let $\{P_\theta, \theta \in \Theta \subseteq \mathbb{R}^p\}$ be of the form

$$\frac{dP_\theta}{dx}(x) := p_\theta(x) = \exp\left(\langle T(x), \Theta \rangle_{\mathbb{R}^p} - G(\theta) + q(x)\right),$$

where $\langle \cdot, \cdot \rangle_{\mathbb{R}^p}$ is the standard Euclidean dot product. Then the Hellinger distance can be expressed as (Nielsen and Garcia, 2011)

$$h^2(P_{\theta_1}, P_{\theta_2}) = 1 - \exp\left(-\frac{1}{2}\left(G(\theta_1) + G(\theta_2) - 2G\left(\frac{\theta_1 + \theta_2}{2}\right)\right)\right).$$

If $G(\theta)$ is convex and its Hessian $D^2G(\theta)$ satisfies $D^2G(\theta) \succeq A$ uniformly for all $\theta \in \Theta$ and some symmetric positive definite operator $A : \mathbb{R}^p \mapsto \mathbb{R}^p$, then

$$h^2(P_{\theta_1}, P_{\theta_2}) \geq 1 - \exp\left(-\frac{1}{8}(\theta_1 - \theta_2)^T A(\theta_1 - \theta_2)\right),$$

hence assumption 3.6 holds with $\tilde{C} = \frac{1}{\sqrt{2}}$ and $\gamma = 1$ for

$$k(\theta_1, \theta_2) := \exp\left(-\frac{1}{8}(\theta_1 - \theta_2)^T A(\theta_1 - \theta_2)\right).$$

Assume that $\Theta \subset \mathbb{R}^p$, let $k(\cdot, \cdot)$ be a symmetric positive definite kernel defined on $\mathbb{R}^p \times \mathbb{R}^p$ and $(\mathbb{H}, \langle \cdot, \cdot \rangle_{\mathbb{H}})$ – a corresponding reproducing kernel Hilbert space (RKHS) with norm $\mathbb{H} \ni f \mapsto \|f\|_{\mathbb{H}} := \sqrt{\langle f, f \rangle_{\mathbb{H}}}$. In what follows, we will assume that k satisfies conditions of proposition 2.3 (in particular, k is characteristic). Often, ρ_k can be bounded above by the Euclidean norm $\|\cdot\|_2$, as shown by the following proposition.

Recall that by Bochner's theorem, there exists a finite nonnegative Borel measure ν such that $k(\theta) = \int_{\mathbb{R}^p} e^{i\langle x, \theta \rangle} d\nu(x)$.

PROPOSITION 3.8. *Assume that $\int_{\mathbb{R}^p} \|x\|_2^2 d\nu(x) < \infty$. Then there exists $D_k > 0$ depending only on k such that for all θ_1, θ_2 ,*

$$\rho_k(\theta_1, \theta_2) \leq D_k \|\theta_1 - \theta_2\|_2.$$

PROOF. For all $z \in \mathbb{R}$, $|e^{iz} - 1 - iz| \leq \frac{|z|^2}{2}$, implying that

$$\begin{aligned} \rho_k^2(\theta_1, \theta_2) &= \|k(\cdot, \theta_1) - k(\cdot, \theta_2)\|_{\mathbb{H}}^2 = 2k(0) - 2k(\theta_1 - \theta_2) = 2 \int_{\mathbb{R}^p} (1 - e^{i\langle x, \theta_1 - \theta_2 \rangle}) d\nu(x) \\ &\leq \int_{\mathbb{R}^p} \langle x, \theta_1 - \theta_2 \rangle_{\mathbb{R}^p}^2 d\nu(x) \leq \|\theta_1 - \theta_2\|_2^2 \int_{\mathbb{R}^p} \|x\|_2^2 d\nu(x). \end{aligned}$$

□

The aforementioned facts easily imply the following statement.

COROLLARY 3.9. *Assume that kernel k satisfies conditions of Proposition 3.8 and that $\Theta \subset \mathbb{R}^p$ is compact. Suppose that for all $\theta_1, \theta_2 \in \Theta$ and some $\gamma > 0$,*

$$(3.8) \quad h(P_{\theta_1}, P_{\theta_2}) \geq c(\Theta) \|\theta_1 - \theta_2\|_2^\gamma.$$

Then assumption 3.6 holds with γ as above and $\tilde{C} = \tilde{C}(k, c(\Theta), \gamma)$.

PROOF. By Proposition 3.8 and (3.8),

$$\rho_k^\gamma(\theta_1, \theta_2) \leq D_k^\gamma \|\theta_1 - \theta_2\|_2^\gamma \leq \frac{D_k^\gamma}{c(\Theta)} h(p_{\theta_1}, p_{\theta_2}).$$

Therefore, $h(P_{\theta_1}, P_{\theta_2}) \geq \frac{c(\Theta)}{D_k^\gamma} \rho_k^\gamma(\theta_1, \theta_2)$.

□

We are ready to state our main result for convergence with respect to the RKHS-induced distance $\|\cdot\|_{\mathcal{F}_k}$.

THEOREM 3.10. *Assume that conditions of Theorem 3.3 hold and that assumption 3.6 is satisfied. Then there exists a sufficiently large $R = R(C, K, \tilde{C}, \gamma) > 0$ such that*

$$(3.9) \quad \Pr \left(d_{W_1, \rho_k}(\delta_0, \Pi_l(\cdot|\mathcal{X}_l)) \geq R\varepsilon_l^{1/\gamma} \right) \leq \frac{1}{l\varepsilon_l^2} + 4e^{-(1+K/2)l\varepsilon_l^2/2},$$

$$(3.10) \quad \Pr \left(\|\delta_0 - \Pi_l(\cdot|\mathcal{X}_l)\|_{\mathcal{F}_k} \geq R\varepsilon_l^{1/\gamma} \right) \leq \frac{1}{l\varepsilon_l^2} + 4e^{-(1+K/2)l\varepsilon_l^2/2}.$$

PROOF. The proof of (3.9) follows immediately from Theorem 3.3 and Assumption 3.6: it is enough to notice that

$$\begin{aligned} d_{W_1, \rho_k}(\delta_0, \Pi_l(\cdot|\mathcal{X}_l)) &= \int_{\Theta} \rho_k(\theta, \theta_0) d\Pi_l(\theta|X_1, \dots, X_l) \leq \\ &R\varepsilon_l^{1/\gamma} + \int_{\rho_k(\theta, \theta_0) \geq R\varepsilon_l^{1/\gamma}} d\Pi_l(\cdot|\mathcal{X}_l) \leq \\ &R\varepsilon_l^{1/\gamma} + 2\|k\|_{\infty} \int_{h(P_{\theta}, P_0) \geq \tilde{C}R^{\gamma}\varepsilon_l} d\Pi_l(\cdot|\mathcal{X}_l), \end{aligned}$$

where $\|k\|_{\infty} := \sup_{\theta \in \Theta} k(\theta, \theta)$; see Appendix B for further details. Finally, (3.10) is a straightforward corollary of (3.9) and the inclusion $\mathcal{F}_k \subseteq \mathcal{F}_L$ (see (3.7)). \square

It is well-known (e.g., Section 5 in Ghosal, Ghosh and Van Der Vaart (2000)) that in the case of finite-dimensional models, requirements of Theorem 3.3 are implied by the inequalities between the Kullback-Leibler, Hellinger and Euclidean distances. This fact, together with Corollary 3.9, implies the following bounds.

THEOREM 3.11. *Suppose that $\Theta \subset \mathbb{R}^p$ is compact. Assume that there exist positive constants $c_i = c_i(\Theta)$, $i = 1, \dots, 4$ such that for all $\theta_1, \theta_2 \in \Theta$*

$$\begin{aligned} P_{\theta_1} \log \frac{p_{\theta_1}}{p_{\theta_2}} &\leq c_1(\Theta) \|\theta_1 - \theta_2\|_2^{2\gamma}, \\ P_{\theta_1} \left(\log \frac{p_{\theta_1}}{p_{\theta_2}} \right)^2 &\leq c_2(\Theta) \|\theta_1 - \theta_2\|_2^{2\gamma}, \\ c_3(\Theta) \|\theta_1 - \theta_2\|_2^{\gamma} &\leq h(P_{\theta_1}, P_{\theta_2}) \leq c_4(\Theta) \|\theta_1 - \theta_2\|_2^{\gamma}. \end{aligned}$$

Assume the prior Π has a density that is uniformly bounded from below. Let $\varepsilon_l = \tau \sqrt{\frac{\log l}{l}}$ for a sufficiently small $\tau > 0$. Then there exists $R = R(c_1, c_2, c_3, c_4, C_k, K)$

$$(3.11) \quad \Pr \left(d_{W_{1, \rho_k}}(\delta_0, \Pi_l(\cdot | \mathcal{X}_l)) \geq R\varepsilon_l^{1/\gamma} \right) \leq \frac{1}{l\varepsilon_l^2} + 4e^{-(1+K/2)l\varepsilon_l^2/2},$$

$$(3.12) \quad \Pr \left(\|\delta_0 - \Pi_l(\cdot | \mathcal{X}_l)\|_{\mathcal{F}_k} \geq R\varepsilon_l^{1/\gamma} \right) \leq \frac{1}{l\varepsilon_l^2} + 4e^{-(1+K/2)l\varepsilon_l^2/2}.$$

PROOF. Let $\Theta_l = \Theta$. Using the inequality between Euclidean and Hellinger distances, it is easy to check that the entropy condition (2) of Theorem 3.3 is satisfied with $\varepsilon_l = \tau \sqrt{\frac{\log l}{l}}$. Also,

$$\begin{aligned} & \Pi \left(P_{\theta_1} \log \frac{p_{\theta_1}}{p_{\theta_2}} \leq \varepsilon_l^2, P_{\theta_1} \left(\log \frac{p_{\theta_1}}{p_{\theta_2}} \right)^2 \leq \varepsilon_l^2 \right) \geq \\ & \Pi \left(\|\theta_1 - \theta_2\|_2^{2\gamma} \leq \frac{\varepsilon_l^2}{\max(c_1(\Theta), c_2(\Theta))} \right) \geq C\varepsilon_l^{p/\gamma}, \end{aligned}$$

where C is a constant depending on $c_1(\Theta)$, $c_2(\Theta)$ and the lower bound on the density of Π . \square

Theorem 3.10 yields the “weak concentration” estimate that is needed to obtain the guarantees for the M-posterior distributions $\hat{\Pi}_{n,g}$ and $\hat{\Pi}_{n,0}$. This is summarized in the following corollary which is one of our main results:

COROLLARY 3.12. *Let X_1, \dots, X_n be an i.i.d. sample from P_0 , and assume that $\hat{\Pi}_{n,g}$ and $\hat{\Pi}_{n,0}$ are defined with respect to the $\|\cdot\|_{\mathcal{F}_k}$ as in (3.3) above. Let $l := \lfloor n/m \rfloor$. Assume that conditions of Theorem 3.10 hold, and, moreover, ε_l is such that $q := \frac{1}{l\varepsilon_l^2} + 4e^{-(1+K/2)l\varepsilon_l^2/2} < \frac{1}{2}$. Let α satisfy $q < \alpha < 1/2$. Then*

$$(3.13) \quad \Pr \left(d_{W_{1, \rho_k}}(\delta_0, \hat{\Pi}_{n,0}) \geq 3R\varepsilon_l^{1/\gamma} \right) \leq e^{-m\psi(1/2, q)}.$$

Let α be such that $q < \alpha < 1/2$. Then

$$(3.14) \quad \Pr \left(\|\delta_0 - \hat{\Pi}_{n,g}\|_{\mathcal{F}_k} \geq C_\alpha R\varepsilon_l^{1/\gamma} \right) \leq e^{-m\psi(\alpha, q)},$$

where $C_\alpha = (1 - \alpha)\sqrt{\frac{1}{1-2\alpha}}$.

PROOF. It is enough to apply parts (a) and (b) of Theorem 2.1 with $\gamma = 0$ to the independent random measures $\Pi_n(\cdot|G_j)$, $j = 1, \dots, m$. Note that the “weak concentration” assumption (2.6) is implied by (3.11) and (3.12). \square

It is also straightforward to see that under the assumptions of Theorem 3.11, (3.13) and (3.16) hold with $\varepsilon_l = \tau \sqrt{\frac{m \log(n/m)}{n}}$ for τ small enough. If $m \simeq \log n$ (which is a typical scenario), then $\varepsilon_l \simeq \sqrt{\frac{\log^2 n}{n}}$. At the same time, if we use the “full posterior” distribution $\Pi_n(\cdot|\mathcal{X}_n)$ (which corresponds to $m = 1$), conclusion of Theorem 3.3 is that

$$\Pr \left(\|\delta_0 - \Pi_n(\cdot|\mathcal{X}_n)\|_{\mathcal{F}_k} \geq R \left(\frac{\log n}{n} \right)^{1/(2\gamma)} \right) \lesssim \log^{-1} n,$$

while Corollary 3.16 yields a much stronger bound for $\hat{\Pi}_{n,g}$:

$$\Pr \left(\|\delta_0 - \hat{\Pi}_{n,g}\|_{\mathcal{F}_k} \geq C_\alpha R \left(\frac{\log^2 n}{n} \right)^{1/(2\gamma)} \right) \leq r_n^{-\log n}$$

for some $r_n \rightarrow 0$.

3.3. Robust Bayesian inference based on stochastic approximations of the posterior distribution. We have seen in the previous section that both $\hat{\Pi}_{n,g}$ and $\hat{\Pi}_{n,0}$ admit strong concentration guarantees. When the number of disjoint subgroups m is large, the resulting M-posterior distribution is very robust, however, at the same time it is often too “flat”, which results in large credible sets and overestimation of uncertainty in applications.

The source of the problem is the fact that each individual random measure $\Pi_n(\cdot|G_j)$, $j = 1, \dots, m$ is based on sample of size $l \simeq \frac{n}{m}$ which can be much smaller than n . The simplest way to “increase” the sample size and to reduce the variance of each subset posterior distribution is to repeat each observation in G_j m times (although other alternatives, such as bootstrap, are possible), $\tilde{G}_j = \underbrace{\{G_j, \dots, G_j\}}_{m \text{ times}}$. Formal application of the Bayes rule in

this situation yields a collection of new measures on the parameter space:

$$\Pi_{n,m}(B|G_j) := \frac{\int_B \left(\prod_{i \in G_j} p_\theta(X_i) \right)^m d\Pi(\theta)}{\int_\Theta \left(\prod_{i \in G_j} p_\theta(X_i) \right)^m d\Pi(\theta)}.$$

Here, $\left(\prod_{i \in G_j} p_\theta(X_i)\right)^m$ can be viewed as an approximation of the full data likelihood. As we have already mentioned in the introduction, random measure $\Pi_{n,m}(\cdot|G_j)$ is called the *j-th stochastic approximation* to the full posterior distribution. Each of the stochastic approximations $\Pi_{n,m}(\cdot|G_j)$ might have unreliable coverage properties, however, their robust version obtained by evaluating the median combines convergence guarantees on the one side with smaller size credible sets (as compared to $\hat{\Pi}_{n,g}$) on the other. Performance of this method in our numerical experiments is particularly impressive.

First, we will show that under certain assumptions the upper bounds for the convergence rates of $\Pi_{n,m}(\cdot|G_j)$ towards δ_0 are the same as for $\Pi_l(\cdot|G_j)$, the “standard posterior distribution” given G_j .

Let $N_{[]}(\mathbf{u}, \Theta, \rho)$ be the bracketing covering number of $\{p_\theta, \theta \in \Theta\}$ with respect to the Hellinger distance ρ , and let $H_{[]}(\mathbf{u}) := \log N_{[]}(\mathbf{u}, \Theta, \rho)$ be the *bracketing entropy*.

THEOREM 3.13 (Wong and Shen (1995), Theorem 1). *There exist constants c_j , $j = 1, \dots, 4$ and $\zeta > 0$ such that if*

$$\int_{\zeta^2/2^8}^{\sqrt{2}\zeta} H_{[]}^{1/2}(u/c_3) du \leq c_4 \sqrt{l} \zeta^2,$$

then

$$P \left(\sup_{\theta: h(P_\theta, P_0) \geq \zeta} \prod_{j=1}^l \frac{p_\theta}{p_0}(X_j) \geq e^{-c_1 l \zeta^2} \right) \leq 4e^{-c_2 l \zeta^2}.$$

In particular, one can choose $c_1 = 1/24$, $c_2 = (4/27)(1/1926)$, $c_3 = 10$ and $c_4 = (2/3)^{5/2}/512$.

In many typical parametric problems, the bracketing entropy can be bounded as $H_{[]}(\mathbf{u}) \lesssim \log(1/u)$, and the minimal ζ that satisfies conditions of Theorem 3.13 is of order $\zeta \simeq \frac{\log l}{\sqrt{l}}$.

Application of the previous theorem to the analysis of “stochastic approximations” yields the following result.

THEOREM 3.14. *Assume that conditions of Theorem 3.13 hold with $\zeta := \varepsilon_l$.*

Moreover, suppose that for some $C > 0$

$$(1) \Pr \left(\theta : -P_0 \left(\log \frac{p_\theta}{p_0} \right) \leq \varepsilon_l^2, P_0 \left(\log \frac{p_\theta}{p_0} \right)^2 \leq \varepsilon_l^2 \right) \geq \exp(-Cl\varepsilon_l^2),$$

$$(2) e^{-ml\varepsilon_l^2} \leq \varepsilon_l.$$

Then there exists $R = R(C) > 0$ such that

$$\Pr \left(d_{W_{1,\rho}}(\delta_0, \Pi_{n,m}(\cdot|\mathcal{X}_l)) \geq (1+R)\varepsilon_l \right) \leq \frac{1}{l\varepsilon_l^2} + 4e^{-c_2 R^2 l \varepsilon_l^2},$$

where ρ is the Hellinger metric and c_2 is a constant from Theorem 3.13.

PROOF. See Appendix C. □

If assumption 3.6 holds, then results of Theorem 3.14 can be easily extended to the bounds on $\|\delta_0 - \Pi_{n,m}(\cdot|\mathcal{X}_l)\|_{\mathcal{F}_k}$.

COROLLARY 3.15. *Assume that conditions of Theorem 3.14 hold and that assumption 3.6 is satisfied for some \tilde{C} and γ . Then there exists $R = R(C, \tilde{C}, \gamma) > 0$ such that*

$$\Pr \left(\|\delta_0 - \Pi_{n,m}(\cdot|\mathcal{X}_l)\|_{\mathcal{F}_k} \geq R\varepsilon_l^{1/\gamma} + \varepsilon_l \right) \leq \frac{1}{l\varepsilon_l^2} + 4e^{-c_2 \tilde{C}^2 R^{2\gamma} l \varepsilon_l^2}.$$

Recall that

$$(3.15) \quad \hat{\Pi}_{n,g}^{\text{st}} := \text{med}_g(\Pi_{n,m}(\cdot|G_1), \dots, \Pi_{n,m}(\cdot|G_m))$$

is the geometric median of $\{\Pi_{n,m}(\cdot|G_1), \dots, \Pi_{n,m}(\cdot|G_m)\}$ with respect to $\|\cdot\|_{\mathcal{F}_k}$. Theorem 2.1 combined with the “weak concentration” inequality of Theorem 3.14 gives the following bound.

COROLLARY 3.16. *Let X_1, \dots, X_n be an i.i.d. sample from P_0 . Let $l := \lfloor n/m \rfloor$. Assume that conditions of Corollary 3.15 hold, and, moreover, ε_l is such that*

$$q := \frac{1}{l\varepsilon_l^2} + 4e^{-c_2 \tilde{C}^2 R^{2\gamma} l \varepsilon_l^2} < \frac{1}{2}.$$

Let α be such that $q < \alpha < 1/2$. Then

$$(3.16) \quad \Pr \left(\|\delta_0 - \hat{\Pi}_{n,g}^{\text{st}}\|_{\mathcal{F}_k} \geq C_\alpha (R\varepsilon_l^{1/\gamma} + \varepsilon_l) \right) \leq e^{-m\psi(\alpha, q)},$$

where $C_\alpha = (1-\alpha)\sqrt{\frac{1}{1-2\alpha}}$.

3.4. M -posterior and model misspecification. In this section, we discuss another type of robustness achieved by our method - namely, robustness with respect to model misspecification.

We will discuss the following example in which the mean θ of the distribution P_θ is our primary interest. Let $X, X_1, \dots, X_n \in \mathbb{R}^p$ be i.i.d from unknown P_0 with mean $\mathbb{E}X = \theta_0$ and covariance matrix $\mathbb{E}(X - \theta_0)(X - \theta_0)^T = \Sigma_0$. In this case, a possible (and probably simplest) model assumption is $P_0 \in \{P_\theta = N(\theta, \sigma^2 I_p), \theta \in \mathbb{R}^p\}$, where I_p is the $p \times p$ identity matrix and $\sigma^2 > 0$ is fixed. Finally, we impose a normal prior $\theta \sim N(0, I_p)$ (which is a conjugate prior for the model). Of course, if the true underlying distribution is not normal but rather heavy-tailed, the usual posterior will be very sensitive to outliers. However, M -posterior is much more robust, as shown in the following statement.

Again, we divide the data X_1, \dots, X_n into $m \leq n/2$ disjoint subgroups G_1, \dots, G_m of size $\geq l := \lfloor n/m \rfloor$ each. The following proposition shows that the M -posterior achieves strong concentration around the true mean parameter θ_0 .

PROPOSITION 3.17. *Set $k(\theta_1, \theta_2) := \exp\left(-\frac{1}{8\sigma^2}(\theta_1 - \theta_2)^T(\theta_1 - \theta_2)\right)$. Let*

$$\hat{\Pi}_{n,g}^{\text{st}} = \text{med}_g(\Pi_{n,m}(\cdot|G_1), \dots, \Pi_{n,m}(\cdot|G_m))$$

be defined with respect to the $\|\cdot\|_{\mathcal{F}_k}$ distance. Then for $s > 2$ and

$$\varepsilon(z, s) := \sqrt{\frac{p\sigma^2}{lm + \sigma^2}}(1 + \sqrt{z/p}) + \sqrt{\frac{s \cdot \text{Tr}(\Sigma_0)}{l}} + \frac{\|\theta_0\|_2}{lm + \sigma^2} + 2\sqrt{2} \exp(-z/2),$$

$$(3.17) \quad \Pr\left(\|\delta_0 - \hat{\Pi}_{n,g}^{\text{st}}\|_{\mathcal{F}_k} \geq C_\alpha \varepsilon(z, s)\right) \leq e^{-m\psi(\alpha, 1/s)},$$

where α is such that $1/s < \alpha < 1/2$ and $C_\alpha = (1 - \alpha)\sqrt{\frac{1}{1-2\alpha}}$.

Note that, for $z \gtrsim \log n$, $\varepsilon(z, s)$ is controlled by $\max\left(\sqrt{\frac{p+\log n}{n}}, \sqrt{\frac{m \text{Tr}(\Sigma_0)}{n}}\right)$. Moreover, (3.17) holds without *any* additional assumptions on P_0 besides finite second moments.

PROOF. We proceed with a direct argument which makes use of the concentration properties of Gaussian measures. Let $\bar{X}_l = \frac{1}{l} \sum_{j=1}^l X_j$ be the

sample mean of the sample $\mathcal{X}_l = \{X_1, \dots, X_l\}$. Since we are in the conjugate prior situation, it is easy to see that the posterior distribution given \mathcal{X}_l is $\Pi_l(\theta|X_1, \dots, X_l) \sim N\left(\frac{l}{l+\sigma^2}\bar{X}_l, \frac{\sigma^2}{l+\sigma^2}I_p\right)$, and the corresponding stochastic approximation is

$$\Pi_{n,m}(\theta|X_1, \dots, X_l) \sim N\left(\frac{lm}{lm+\sigma^2}\bar{X}_l, \frac{\sigma^2}{lm+\sigma^2}I_p\right).$$

Note that, by Chebyshev's inequality, for any $s > 0$,

$$\Pr\left(\|\bar{X}_l - \theta_0\|_2 \geq \sqrt{\frac{s \cdot \text{Tr}(\Sigma_0)}{l}}\right) \leq \frac{1}{s}.$$

Moreover, the concentration inequality for Gaussian measures ([Ledoux, 2001](#)) implies that for all $t \geq 0$,

$$\Pi_{n,m}\left(\theta : \left\|\frac{lm}{lm+\sigma^2}\bar{X}_l - \theta\right\|_2 \geq \sqrt{\frac{p\sigma^2}{lm+\sigma^2}}(1 + \sqrt{t/p}) \middle| X_1, \dots, X_l\right) \leq 2 \exp\left(-\frac{t}{2}\right).$$

Let A_s be an event of probability $\geq 1 - 1/s$ on which $\|\bar{X}_l - \theta_0\|_2 \leq \sqrt{\frac{s \cdot \text{Tr}(\Sigma_0)}{l}}$. On A_s , we have that

$$\begin{aligned} \Pi_{n,m}\left(\theta : \|\theta - \theta_0\|_2 \geq t \middle| X_1, \dots, X_l\right) &\leq \\ \Pi_{n,m}\left(\theta : \left\|\theta - \frac{lm}{lm+\sigma^2}\bar{X}_l\right\|_2 \geq t - \sqrt{\frac{s \cdot \text{Tr}(\Sigma_0)}{l}} - \frac{\|\theta_0\|_2}{lm+\sigma^2} \middle| X_1, \dots, X_l\right) \end{aligned}$$

which is further bounded by $2 \exp(-z/2)$ for

$$t = t(z, s) := \sqrt{\frac{p\sigma^2}{lm+\sigma^2}}(1 + \sqrt{z/p}) + \sqrt{\frac{s \cdot \text{Tr}(\Sigma_0)}{l}} + \frac{\|\theta_0\|_2}{lm+\sigma^2}.$$

Therefore, for each subset G_j , $j = 1, \dots, m$, we have that for all $s > 0$

$$(3.18) \quad \Pr\left(\Pi_{n,m}\left(\theta : \|\theta - \theta_0\|_2 \geq t(s, z) \middle| G_j\right) \leq 2 \exp(-z/2)\right) \geq 1 - 1/s.$$

By the definition of k and the inequality $1 - e^{-z} \leq z$ for $z \in \mathbb{R}$,

$$\rho_k^2(\theta_1, \theta_2) = 2\left(1 - \exp\left(-\frac{1}{8\sigma^2}(\theta_1 - \theta_2)^T(\theta_1 - \theta_2)\right)\right) \leq \frac{1}{4\sigma^2}\|\theta_1 - \theta_2\|_2^2.$$

Together with (3.7), it implies that

$$\begin{aligned} \|\delta_0 - \Pi_{n,m}(\cdot|G_j)\|_{\mathcal{F}_k} &\leq W_{1,\rho_k}(\delta_0, \Pi_{n,m}(\cdot|G_j)) = \int_{\mathbb{R}^p} \rho_k(\theta_0, \theta) d\Pi_{n,m}(\cdot|G_j) \leq \\ &2\sigma t(z, s) + \sqrt{2}\Pi_{n,m}(\theta : \rho_k(\theta_0, \theta) \geq 2\sigma t(s, z)|G_j) \leq \\ &2\sigma t(z, s) + \sqrt{2}\Pi_{n,m}(\theta : \|\theta_0 - \theta\|_2 \geq t(s, z)|G_j). \end{aligned}$$

From (3.18), we conclude that, with probability $\geq 1 - 1/s$,

$$\|\delta_0 - \Pi_{n,m}(\cdot|G_j)\|_{\mathcal{F}_k} \leq 2\sigma t(z, s) + 2\sqrt{2}\exp(-z/2).$$

Combination of this “weak concentration” bound with Theorem 2.1 yields the result. \square

4. Numerical algorithms and examples. In this section, we consider examples and applications in which comparisons are made for the inference based on the usual posterior distribution and on the M-posterior. One of the well-known and computationally efficient ways to find the geometric median in Hilbert spaces is the famous *Weiszfeld’s algorithm* (introduced in Weiszfeld (1936)). Details of implementation are described in Algorithms 1 and 2. Algorithm 1 is a particular case of Weiszfeld’s algorithm applied to subset posterior distributions and distance $\|\cdot\|_{\mathcal{F}_k}$, while Algorithm 2 shows how to obtain an approximation to M-posterior given the samples from $\Pi_{n,m}(\cdot|G_j)$, $j = 1 \dots m$. Note that the subset posteriors $\Pi_{n,m}(\cdot|G_j)$ whose “weights” $w_{*,j}$ in the expression of the M-posterior are small (in our case, smaller than $1/(2m)$) are excluded from the analysis. Our extensive simulations show the empirical evidence in favor of this additional thresholding step.

Detailed discussion of convergence rates and acceleration techniques for Weiszfeld’s method from the viewpoint of modern optimization can be found in Beck and Sabach (2013). For alternative approaches and extensions of Weiszfeld’s algorithm, see Bose, Maheshwari and Morin (2003), Ostresh (1978), Overton (1983), Chandrasekaran and Tamir (1990), Cardot, Cénac and Zitt (2012), Cardot, Cénac and Zitt (2013), among other works.

Before presenting the results of numerical analysis, let us remark on two important computational aspects not discussed previously.

REMARK 4.1.

Algorithm 1 Evaluating the geometric median of probability distributions via Weiszfeld’s algorithm

Input:

1. Discrete measures Q_1, \dots, Q_m ;
2. The kernel $k(\cdot, \cdot) : \mathbb{R}^p \times \mathbb{R}^p \mapsto \mathbb{R}$;
3. Threshold $\varepsilon > 0$;

Initialize:

1. Set $w_j^{(0)} := \frac{1}{m}$, $j = 1 \dots m$;
2. Set $Q_*^{(0)} := \frac{1}{m} \sum_{j=1}^m Q_j$;

repeat

Starting from $t = 0$, for each $j = 1, \dots, m$:

1. Update $w_j^{(t+1)} = \frac{\|Q_*^{(t)} - Q_j\|_{\mathcal{F}_k}^{-1}}{\sum_{i=1}^m \|Q_*^{(t)} - Q_i\|_{\mathcal{F}_k}^{-1}}$; (apply (2.12) to evaluate $\|Q_*^{(t)} - Q_i\|_{\mathcal{F}_k}$);
2. Update $Q_*^{(t+1)} = \sum_{j=1}^m w_j^{(t+1)} Q_j$;

until $\|Q_*^{(t+1)} - Q_*^{(t)}\|_{\mathcal{F}_k} \leq \varepsilon$

Return: $w_* := (w_1^{(t+1)}, \dots, w_m^{(t+1)})$.

(a) The first comment is related to the choice of m , the number of data subsets. So far, it has only been a “free parameter” that enters the theoretical guarantees for our method. However, in many applications we would like to be able to choose an “optimal” m from an interval $[m_1, m_2]$ of acceptable values that is usually dictated by the sample size and computational resources (e.g., the number of available machines). Large values of m provide more robustness but might yield to overestimated uncertainty. We suggest the following heuristic approach which picks the median among the candidate M -posteriors. Namely, start by evaluating the M -posterior for each m in the range $[m_1, m_2]$:

$$\begin{aligned} \hat{\Pi}_{n,m_1}^g &:= \text{med}_g(\Pi_n^{(1)}, \dots, \Pi_n^{(m_1)}), \\ \hat{\Pi}_{n,m_1+1}^g &:= \text{med}_g(\Pi_n^{(1)}, \dots, \Pi_n^{(m_1+1)}), \\ &\quad \dots \dots \dots \\ \hat{\Pi}_{n,m_2}^g &:= \text{med}_g(\Pi_n^{(1)}, \dots, \Pi_n^{(m_2)}) \end{aligned}$$

and choose $m_* \in [m_1, m_2]$ such that

$$(4.1) \quad \hat{\Pi}_{n,m_*}^g = \text{med}_0 \left(\hat{\Pi}_{n,m_1}^g, \hat{\Pi}_{n,m_1+1}^g, \dots, \hat{\Pi}_{n,m_2}^g \right),$$

Algorithm 2 Approximating the M-posterior distribution

Input:

1. Samples $\{Z_{j,i}\}_{i=1}^{N_j} \sim \Pi_{n,m}(\cdot|G_j)$, $j = 1 \dots m$ (see equation (3.4));

Do:

1. $Q_j := \frac{1}{N_j} \sum_{i=1}^{N_j} \delta_{Z_{j,i}}$, $j = 1 \dots m$ - empirical approximations of $\Pi_{n,m}(\cdot|G_j)$.
2. Apply Algorithm 1 to Q_1, \dots, Q_m ; return $w_* = (w_{*,1} \dots w_{*,m})$;
3. For $j = 1, \dots, m$, set $\bar{w}_j := w_{*,j} I\{w_{*,j} \geq \frac{1}{2m}\}$; define $\hat{w}_j^* := \bar{w}_j / \sum_{i=1}^m \bar{w}_i$.

Return: $\hat{\Pi}_{n,g}^{\text{st}} := \sum_{i=1}^m \hat{w}_i^* Q_i$.

where med_0 is the *metric median* defined in (2.3). Figure 5 illustrates this heuristic approach.

(b) Second, it is easy to estimate the improvement in computational time complexity achieved by M-posterior. Given the data set $\mathcal{X}_n = \{X_1, \dots, X_n\}$ of size n , let $t(n)$ be the running time of the algorithm (e.g., MCMC) that outputs a single observation from the posterior distribution $\Pi_n(\cdot|\mathcal{X}_n)$. If the goal is to obtain a sample of size N from the posterior, then the total running time is $O(N \cdot t(n))$. Let us compare this time with the running time needed to obtain a sample of same size N (assuming that N is large) from the M -posterior given that the algorithm is running on m machines ($m \ll n$) in parallel. In this case, we need to generate $O(N/m)$ samples from each of m subset posteriors, which is done in time $O(\frac{N}{m} \cdot t(\frac{n}{m}))$. According to Theorem 7.1 in Beck and Sabach (2013), Weiszfeld's algorithm approximates the M-posterior to degree of accuracy ε in at most $O(1/\varepsilon)$ steps, and each of these steps has complexity $O(N^2)$ (which follows from (2.12)), so that the total running time is

$$O\left(\frac{N}{m} \cdot t\left(\frac{N}{m}\right) + \frac{N^2}{\varepsilon}\right).$$

If, for example, $t(n) \simeq n^r$ for some $r \geq 1$, then $\frac{N}{m} \cdot t\left(\frac{n}{m}\right) \simeq \frac{1}{m^{1+r}} N n^r$ which should be compared to $N \cdot n^r$ required by the standard approach. The term $\frac{N^2}{\varepsilon}$ can be refined in several ways via application of more advanced optimization techniques (see the aforementioned references).

4.1. Numerical analysis: simulated data. In this subsection, we present two examples illustrating robustness and improvements in computational complexity achieved by our method.

4.1.1. *Univariate Gaussian data.* The goal of this example is to demonstrate the effect of the magnitude of an outlier on the posterior distribution of the mean parameter μ . To this end, we used the simplest univariate Gaussian model $\{P_\mu = \mathcal{N}(\mu, 1), \mu \in \mathbb{R}\}$.

We simulated 25 sets containing 100 observations each. Each sample $\{\mathbf{x}_i\}_{i=1}^{25}$, where $\mathbf{x}_i = (x_{i,1}, \dots, x_{i,100})$, contains 99 independent observations from the standard Gaussian distribution ($x_{i,j} \sim \mathcal{N}(0, 1)$ for $i = 1, \dots, 25$ and $j = 1, \dots, 99$), while the last entry in each sample, $x_{i,100}$, is an outlier, and its value linearly increases for $i = 1, \dots, 25$, namely, $x_{i,100} = i \max(|x_{i,1}|, \dots, |x_{i,99}|)$. Index of an outlier is assumed to be unknown to the algorithm, and the variance of observations is known. We generated 50 replications of such data sets. We use a flat (Jeffreys) prior on the mean μ and obtain its posterior distribution, which is also Gaussian with mean $\frac{\sum_{j=1}^{100} x_{ij}}{100}$ and variance $\frac{1}{100}$ (e.g., see Gelman et al. (2003)). We generate 1000 samples from each posterior distribution $\Pi_{100}(\cdot | \mathbf{x}_i)$ for $i = 1, \dots, 25$. Algorithm 2 generates 1000 samples from the M-posterior $\hat{\Pi}_{100,g}^{\text{st}}(\cdot | \mathbf{x}_i)$ for each $i = 1, \dots, 25$: to this end, we set $m = 10$ and generate 100 samples from every $\Pi_{100,10}(\cdot | G_{j,i})$, $j = 1, \dots, 10$ to form the empirical measures $Q_{j,i}$; here, $\cup_{j=1}^{10} G_{j,i} = \mathbf{x}_i$. This process is repeated for all the 50 replications of simulated data. We used Consensus MCMC (introduced in (Scott et al., 2013)) as a representative for scalable MCMC methods, and compared its performance with M-posterior when the number of data subsets is fixed.

Figure 1 compares the performance of the “consensus posterior”, the overall posterior and the M-posterior using the empirical coverage of $(1-\alpha)100\%$ credible intervals (CIs) calculated across 50 replications for $\alpha = 0.2, 0.15, 0.10$, and 0.05. The empirical coverages of M-posterior’s CIs show robustness to the size of an outlier. On the contrary, performance of the consensus and overall posteriors deteriorate fairly quickly across all α ’s leading to 0% empirical coverage as the outlier strength increases from $i = 1$ to $i = 25$. Figure 3 plots the relative lengths of CIs for the overall posterior and median posterior, with a zero value corresponding to identical lengths and a positive value to wider median posterior intervals. We find that overall the interval widths are similar, but the M-posterior intervals are slightly wider in the absence of large outliers.

4.1.2. *Gaussian process regression.* We use function $f_0(x) = 1 + 3\sin(2\pi x - \pi)$ and simulate 90 (case 1) and 980 (case 2) values of f_0 at equidistant x ’s in $[0, 1]$ (hereafter $x_{1:90}$ and $x_{1:980}$) corrupted by Gaussian noise with mean 0 and variance 1. To demonstrate the robustness of M-posterior in

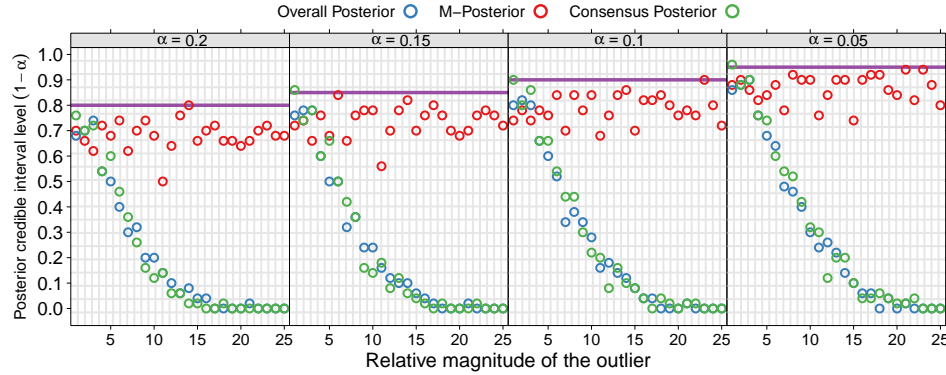


Fig 1: Effect of outlier on the empirical coverage of $(1-\alpha)100\%$ credible intervals (CIs). The x-axis represents the outlier magnitude. The y-axis represents the fraction of times the CIs include the true mean over 50 replications. The panels show the coverage results when $\alpha = 0.2, 0.15, 0.10$, and 0.05 . The horizontal lines (in violet) show the frequentist coverage.

nonparametric regression, we added 10 (case 1) and 20 (case 2) outliers (sampled on the uniform grids of corresponding sizes) to the data sets such that $f_0(x_{91:100}) = 10 \max(f_0(x_{1:90}))$ and $f_0(x_{981:1000}) = 10 \max(f_0(x_{1:980}))$.

The `gausspr` function in `kernlab` R package (Karatzoglou et al., 2004) is used for GP regression. Based on the standard convention in GP regression, the noise variance (or “nugget effect”) is fixed at 0.01. Using these settings for GP regression without the “standard” posterior, `gausspr` obtains an estimator \hat{f}_1 and a 95% confidence band for the values of the regression function at 100 equally spaced grid points $y_{1:100}$ in $[0, 1]$ (note that these locations are different from the observed data). Algorithm 2 performs GP regression with M-posterior and obtains an estimator \hat{f}_2 described below. The posterior draws across $y_{1:100}$ are obtained in cases 1 and 2 as follows. First, $\{(x_i, f_i)\}$ are split into $m = 10$ and 20 subsets (each living on its own uniform grid) respectively, and `gausspr` estimates the posterior mean μ_j and covariance Σ_j for each data subset, $j = 1, \dots, m$. These estimates correspond to the Gaussian distributions $\Pi_j(\cdot | \mu_j, \Sigma_j)$ that are used to generate 100 posterior draws at $y_{1:100}$ each. These draws are further employed to form the empirical versions of subset posteriors. Finally, Weiszfeld’s algorithm is used to combine them. Next, we obtained 1000 samples from the M-posterior $\Pi_g(\cdot | \{(x_i, f_i)\})$. The median of these 1000 samples at each location on the grid $y_{1:100}$ represents the estimator \hat{f}_2 . Its 95% confidence band corresponds

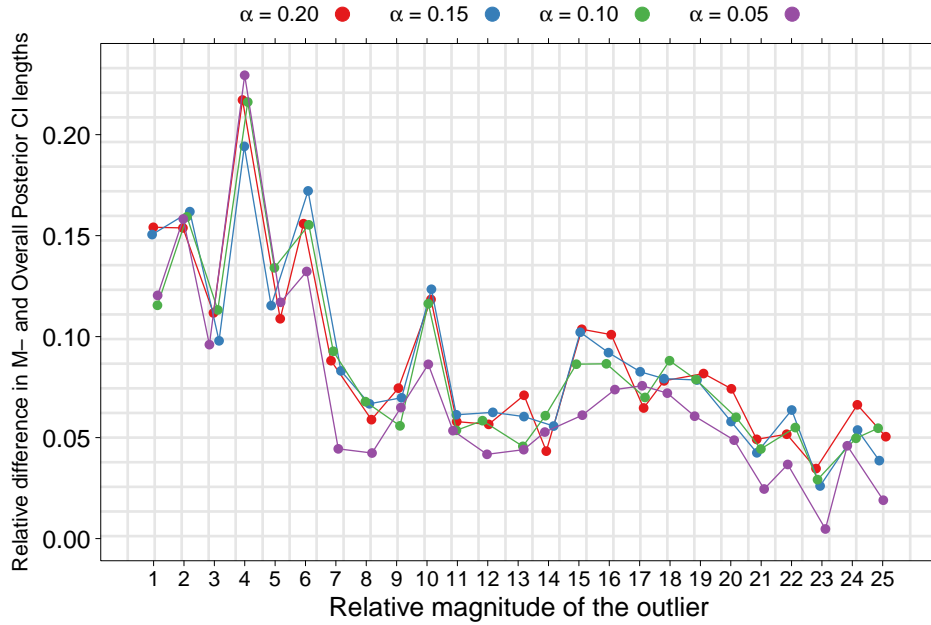


Fig 2: Calibration of uncertainty quantification of M-posterior. The x-axis represents the outlier magnitude that increases from 1 to 25. The y-axis represents ratio of difference in the CI lengths of M-posterior. The different colored lines represent calibration when $\alpha = 0.2, 0.15, 0.10$, and 0.05 . A value close to 0 represents that the M-posterior CIs are well-calibrated.

to 2.5% and 97.5% quantiles of the 1000 posterior draws across $y_{1:100}$.

Figure 3 summarizes the results of GP regression with and without M-posterior across 30 replications. In case 1, GP regression without M-posterior is extremely sensitive to the outliers, resulting in \hat{f}_1 that is shifted above the truth and distorted near the x 's that are adjacent to the outliers; in turn, this affects the coverage of 95% confidence bands and results in the “bumps” that correspond to the location of outliers. In contrast, GP regression using M-posterior produces \hat{f}_2 which is close to the true curve in both cases; however, in case 1, when the number of data points is small, the 95% bands are unstable.

An attractive property of M-posterior based GP regression is that numerical instability due to matrix inversion can be avoided by working with multiple subsets. We investigated such cases when the number of data points n was

greater than 10^4 . Chalupka, Williams and Murray (2012) compare several low rank matrix approximations techniques used to avoid matrix inversion in massive data GP computation. M-posterior-based GP computation does not use approximations to obtain subset posteriors. By increasing the number of subsets (m), M-posterior based GP regression is both computationally feasible and numerically stable for cases when $n = \mathcal{O}(10^6)$ and $m = \mathcal{O}(10^3)$. On the contrary, standard GP regression using the whole data set was intractable for data size greater than 10^4 due of numerical instabilities in matrix inversion. In general, for n data points and m subsets, the computational complexity for GP with M-posterior is $\mathcal{O}(n(\frac{n}{m})^2)$; therefore, $m > 1$ is computationally better than working with the whole data set. By carefully choosing the $\frac{n}{m}$ ratio depending on the available computational resources and n , GP regression with M-posterior is a promising approach for GP regression for massive data without low rank approximations.

4.2. Numerical analysis: real data. The General Social Survey (GSS; <http://www3.norc.oregonstate.edu/gss>) polls Americans for views on different issues, including support for abortion (Abort), capital punishment (Cap), and legalization of marijuana (Mar). These three questions have two possible answers: *yes* or *no*. We use GSS data from 2008 and 2010, consisting of 4067 responders, with about 27% of the data missing and possibilities of contamination (e.g., a small number of survey respondents purposely answering questions incorrectly). We use a Dirichlet process (DP) mixture of product multinomial distributions, probabilistic parafac (p-parafac), to model the multivariate dependence in these data (Dunson and Xing, 2009). We represent the probability for a response Abort = a , Mar = r , and Cap = c as π_{arc} , where $a \in \{\text{yes}, \text{no}\}$, $r \in \{\text{yes}, \text{no}\}$ and $c \in \{\text{yes}, \text{no}\}$. Detailed description of the p-parafac generative model is included in Appendix D. For computational efficiency and robustness, we randomly divided the GSS data into $m = 30$ subsets G_j , $j = 1, \dots, 30$. Using a modified form of the Gibbs sampler that accounts for stochastic approximation, we generated 200 posterior draws from $\hat{\Pi}_{4067,g}^{\text{st}}(\cdot|G_j)$ for $j = 1, \dots, 30$. Using the Weiszfeld's algorithm, we estimated the M-posterior and removed the atoms with estimated weights below $1/60 (= 1/(2m))$.

Figure 4 shows the M-posterior and associated subset posteriors for marginal probabilities of Abort ($\pi_{\text{no}\bullet\bullet}$), Mar ($\pi_{\bullet\text{no}\bullet}$), and Cap ($\pi_{\bullet\bullet\text{yes}}$). All the M-posteriors have their modes close to the maximum likelihood (ML) estimate obtained from the whole data. Figure 5 shows the effect of number of subsets (m) on the M-posterior for the three categories of GSS data. The mode is

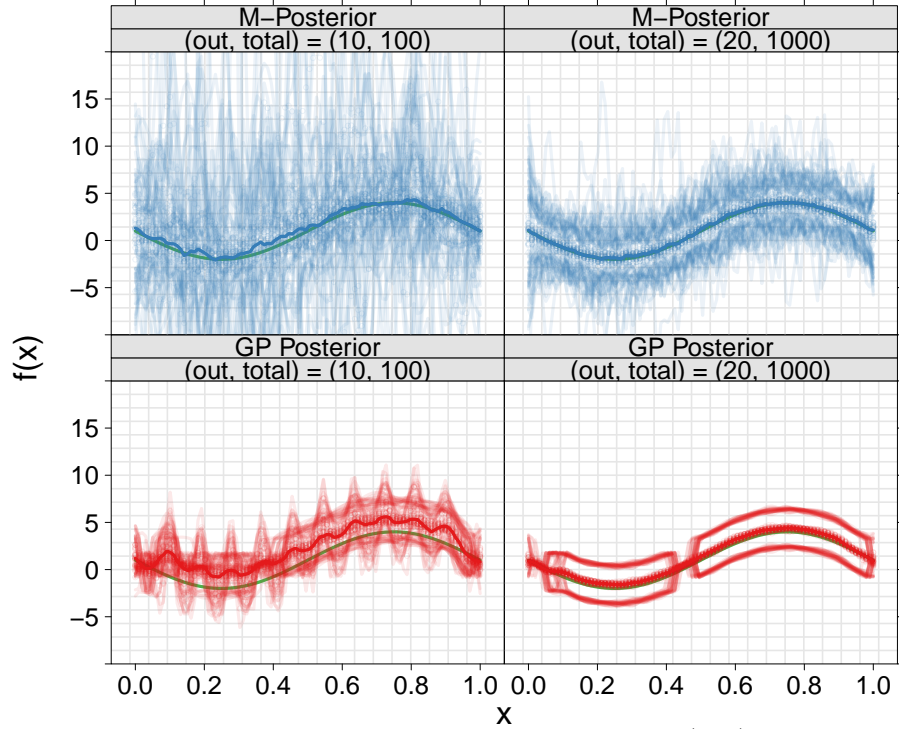


Fig 3: Performance of M-posterior in Gaussian process (GP) regression. The top and bottom row of panels show simulation results for M-posterior (in blue) and GP regression (in red). The size of data increases from column 1 to 2. The true noiseless curve $f_0(x)$ is in green. The shaded regions around the curves represent 95% confidence bands obtained over 30 replicated data sets.

similar in each case, but the variance tends to be smaller when the number of subsets is small, stabilizing as m increases. When m is very small (e.g., 5) the M-posterior is similar to the full data posterior, and can only accommodate a few outliers at most, while as m increases, robustness is improved. This can potentially explain the slight variance inflation. As kernel smoothing based on a modest number of posterior draws was used in estimating these densities, there is no significant difference for $m \geq 25$.

References.

- AGARWAL, A. and DUCHI, J. C. (2012). Distributed delayed stochastic optimization. In *2012 IEEE 51st Annual Conference on Decision and Control (CDC)* 5451–5452.
- AHN, S., KORATTIKARA, A. and WELLING, M. (2012). Bayesian posterior sampling via stochastic gradient Fisher scoring. *Proceedings of the 29th International Conference on Machine Learning (ICML-12)*.
- ALON, N., MATIAS, Y. and SZEGEDY, M. (1996). The space complexity of approximating the frequency moments. In *Proceedings of the twenty-eighth annual ACM symposium on Theory of computing* 20–29. ACM.
- ARONSAJN, N. (1950). Theory of reproducing kernels. *Transactions of the American mathematical society* **68** 337–404.
- BAYARRI, M. and BERGER, J. O. (1994). Robust Bayesian bounds for outlier detection. *Recent Advances in Statistics and Probability* 175–190.
- BECK, A. and SABACH, S. (2013). Weiszfeld’s method: old and new results. *Preprint*. Available at https://iew3.technion.ac.il/Home/Users/becka/Weiszfeld_review-v3.pdf.
- BERGER, J. O. (1994). An overview of robust Bayesian analysis. *Test* **3** 5–124. With comments and a rejoinder by the author.
- BISHOP, C. M. (2006). *Pattern recognition and machine learning*. Springer, New York.
- BOSE, P., MAHESHWARI, A. and MORIN, P. (2003). Fast approximations for sums of distances, clustering, and the Fermat–Weber problem. *Computational Geometry* **24** 135–146.
- BOX, G. E. and TIAO, G. C. (1968). A Bayesian approach to some outlier problems. *Biometrika* **55** 119–129.
- BOYD, S., PARIKH, N., CHU, E., PELEATO, B. and ECKSTEIN, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine Learning* **3** 1–122.
- BRODERICK, T., BOYD, N., WIBISONO, A., WILSON, A. C. and JORDAN, M. (2013). Streaming variational Bayes. In *Advances in Neural Information Processing Systems* 1727–1735.
- CARDOT, H., CÉNAC, P. and ZITT, P.-A. (2012). Recursive estimation of the conditional geometric median in Hilbert spaces. *Electronic Journal of Statistics* **6** 2535–2562.
- CARDOT, H., CÉNAC, P. and ZITT, P.-A. (2013). Efficient and fast estimation of the geometric median in Hilbert spaces with an averaged stochastic gradient algorithm. *Bernoulli* **19** 18–43.
- CHALUPKA, K., WILLIAMS, C. K. and MURRAY, I. (2012). A Framework for Evaluating Approximation Methods for Gaussian Process Regression. *arXiv preprint arXiv:1205.6326*.

- CHANDRASEKARAN, R. and TAMIR, A. (1990). Algebraic optimization: the Fermat-Weber location problem. *Mathematical Programming* **46** 219–224.
- DUDLEY, R. M. (2002). *Real analysis and probability* **74**. Cambridge University Press.
- DUNSON, D. B. and XING, C. (2009). Nonparametric Bayes modeling of multivariate categorical data. *Journal of the American Statistical Association* **104** 1042–1051.
- GELMAN, A., CARLIN, J. B., STERN, H. S. and RUBIN, D. B. (2003). *Bayesian Data Analysis*, 2 ed. Chapman & Hall/CRC, Boca Raton, Florida.
- GHOSAL, S., GHOSH, J. K. and VAN DER VAART, A. W. (2000). Convergence rates of posterior distributions. *Annals of Statistics* **28** 500–531.
- HOFFMAN, M. D., BLEI, D. M., WANG, C. and PAISLEY, J. (2013). Stochastic Variational Inference. *Journal of Machine Learning Research* **14** 1303–1347.
- HSU, D. and SABATO, S. (2013). Loss minimization and parameter estimation with heavy tails. *arXiv preprint arXiv:1307.1827*.
- HUBER, P. J. and RONCHETTI, E. M. (2009). *Robust statistics*, second ed. *Wiley Series in Probability and Statistics*. John Wiley & Sons Inc.
- ISHWARAN, H. and JAMES, L. F. (2001). Gibbs sampling methods for stick-breaking priors. *Journal of the American Statistical Association* **96**.
- KARATZOGLOU, A., SMOLA, A., HORNIK, K. and ZEILEIS, A. (2004). kernlab – An S4 Package for Kernel Methods in R. *Journal of Statistical Software* **11** 1–20.
- KEMPERMAN, J. H. B. (1987). The median of a finite measure on a Banach space. *Statistical Data Analysis Based on the L_1 -norm and Related Methods, North-Holland, Amsterdam* 217–230.
- KORATTIKARA, A., CHEN, Y. and WELLING, M. (2013). Austerity in MCMC land: cutting the Metropolis-Hastings budget. *arXiv preprint arXiv:1304.5299*.
- LEDoux, M. (2001). *The concentration of measure phenomenon. Mathematical Surveys and Monographs* **89**. American Mathematical Society, Providence, RI.
- LERASLE, M. and OLIVEIRA, R. I. (2011). Robust empirical mean estimators. *arXiv preprint arXiv:1112.3914*.
- LOPUHAA, H. P. and ROUSSEEUW, P. J. (1991). Breakdown points of affine equivariant estimators of multivariate location and covariance matrices. *The Annals of Statistics* 229–248.
- MINSKER, S. (2013). Geometric median and robust estimation in Banach spaces. *arXiv preprint arXiv:1308.1334*.
- NEISWANGER, W., WANG, C. and XING, E. (2013). Asymptotically exact, embarrassingly parallel MCMC. *arXiv preprint arXiv:1311.4780*.
- NEMIROVSKI, A. and YUDIN, D. (1983). Problem complexity and method efficiency in optimization.
- NIELSEN, F. and GARCIA, V. (2011). Statistical exponential families: a digest with flash cards. *arXiv preprint arXiv:0911.4863*.
- OSTRESH, L. M. (1978). On the convergence of a class of iterative methods for solving the Weber location problem. *Operations Research* **26** 597–609.
- OVERTON, M. L. (1983). A quadratically convergent method for minimizing a sum of Euclidean norms. *Mathematical Programming* **27** 34–63.
- SCOTT, S. L., BLOCKER, A. W., BONASSI, F. V., CHIPMAN, H. A., GEORGE, E. I. and McCULLOCH, R. E. (2013). Bayes and big data: the consensus Monte Carlo algorithm.
- SETHURAMAN, J. (1994). A constructive definition of Dirichlet measures. *Statistica Sinica* **4** 639–650.
- SMOLA, A. J. and NARAYANAMURTHY, S. (2010). An architecture for parallel topic models. In *Very Large Databases (VLDB)*.
- SRIPERUMBUDUR, B. K., FUKUMIZU, K., GRETTON, A., SCHÖLKOPF, B. and LANCK-

- RIET, G. R. (2009). On integral probability metrics, ϕ -divergences and binary classification. *arXiv preprint arXiv:0901.2698*.
- SRIPERUMBUDUR, B. K., GRETTON, A., FUKUMIZU, K., SCHÖLKOPF, B. and LANCKRIET, G. R. (2010). Hilbert space embeddings and metrics on probability measures. *The Journal of Machine Learning Research* **99** 1517–1561.
- SVENSEN, M. and BISHOP, C. M. (2005). Robust Bayesian mixture modelling. *Neurocomputing* **64** 235–252.
- VAN DER VAART, A. W. and WELLNER, J. A. (1996). *Weak convergence and empirical processes. Springer Series in Statistics*. Springer-Verlag, New York.
- WANG, X. and DUNSON, D. B. (2013). Parallel MCMC via Weierstrass sampler. *arXiv preprint arXiv:1312.4605*.
- WANG, C., PAISLEY, J. W. and BLEI, D. M. (2011). Online variational inference for the hierarchical Dirichlet process. In *International Conference on Artificial Intelligence and Statistics* 752–760.
- WEISZFELD, E. (1936). Sur un problème de minimum dans l’espace. *Tohoku Mathematical Journal*.
- WELLING, M. and TEH, Y. W. (2011). Bayesian learning via stochastic gradient Langevin dynamics. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)* 681–688.
- WONG, W. H. and SHEN, X. (1995). Probability inequalities for likelihood ratios and convergence rates of sieve MLEs. *The Annals of Statistics* 339–362.

APPENDIX A: PROOF OF THEOREM 2.1

We start by proving part **a**. To this end, we will need the following lemma (see lemma 2.1 in [Minsker \(2013\)](#)):

LEMMA A.1. *Let \mathbb{H} be a Hilbert space, $x_1, \dots, x_m \in \mathbb{H}$ and let x_* be their geometric median. Fix $\alpha \in (0, \frac{1}{2})$ and assume that $z \in \mathbb{H}$ is such that $\|x_* - z\| > C_\alpha r$, where*

$$C_\alpha = (1 - \alpha) \sqrt{\frac{1}{1 - 2\alpha}}$$

and $r > 0$. Then there exists a subset $J \subseteq \{1, \dots, m\}$ of cardinality $|J| > \alpha m$ such that for all $j \in J$, $\|x_j - z\| > r$.

Assume that event $\mathcal{E} := \left\{ \|\hat{\theta}_* - \theta_0\| > C_\alpha \varepsilon \right\}$ occurs. Lemma [A.1](#) implies that there exists a subset $J \subseteq \{1, \dots, m\}$ of cardinality $|J| \geq \alpha k$ such that

$\|\hat{\theta}_j - \theta_0\| > \varepsilon$ for all $j \in J$, hence

$$\begin{aligned} \Pr(\mathcal{E}) &\leq \Pr\left(\sum_{j=1}^m I\left\{\|\hat{\theta}_j - \theta_0\| > \varepsilon\right\} > \alpha m\right) \leq \\ &\Pr\left(\sum_{j=1}^{\lfloor (1-\gamma)m \rfloor + 1} I\left\{\|\hat{\theta}_j - \theta_0\| > \varepsilon\right\} > (\alpha - \gamma)m \frac{\lfloor (1-\gamma)m \rfloor + 1}{\lfloor (1-\gamma)m \rfloor + 1}\right) \leq \\ &\Pr\left(\sum_{j=1}^{\lfloor (1-\gamma)m \rfloor + 1} I\left\{\|\hat{\theta}_j - \theta_0\| > \varepsilon\right\} > \frac{\alpha - \gamma}{1 - \gamma}(\lfloor (1-\gamma)m \rfloor + 1)\right). \end{aligned}$$

If W has Binomial distribution $W \sim B(\lfloor (1-\gamma)m \rfloor + 1, q)$, then

$$\begin{aligned} \Pr\left(\sum_{j=1}^{\lfloor (1-\gamma)m \rfloor + 1} I\left\{\|\hat{\theta}_j - \theta_0\| > \varepsilon\right\} > \frac{\alpha - \gamma}{1 - \gamma}(\lfloor (1-\gamma)m \rfloor + 1)\right) \leq \\ \Pr\left(W > \frac{\alpha - \gamma}{1 - \gamma}(\lfloor (1-\gamma)m \rfloor + 1)\right) \end{aligned}$$

(see Lemma 23 in [Lerasle and Oliveira \(2011\)](#) for a rigorous proof of this fact). Chernoff bound (e.g., Proposition A.6.1 in [van der Vaart and Wellner \(1996\)](#)), together with an obvious bound $\lfloor (1-\gamma)m \rfloor + 1 > (1-\gamma)m$, implies that

$$\Pr\left(W > \frac{\alpha - \gamma}{1 - \gamma}(\lfloor (1-\gamma)m \rfloor + 1)\right) \leq \exp\left(-m(1-\gamma)\psi\left(\frac{\alpha - \gamma}{1 - \gamma}, q\right)\right).$$

To establish part **b**, we proceed as follows: let \mathcal{E}_1 be the event

$$\mathcal{E}_1 = \{\text{more than a half of events } d(\hat{\theta}_j, \theta_0) \leq \varepsilon, \ j = 1 \dots m \text{ occur}\}.$$

Assume that \mathcal{E}_1 occurs. Then we clearly have $\varepsilon_* \leq \varepsilon$, where ε_* is defined in (2.2): indeed, for any $\theta_{j_1}, \theta_{j_2}$ such that $d(\hat{\theta}_{j_i}, \theta_0) \leq \varepsilon$, $i = 1, 2$, triangle inequality gives $d(\theta_{j_1}, \theta_{j_2}) \leq 2\varepsilon$. By the definition of $\hat{\theta}_*$, inequality $d(\hat{\theta}_*, \hat{\theta}_j) \leq 2\varepsilon_* \leq 2\varepsilon$ holds for at least a half of $\{\hat{\theta}_1, \dots, \hat{\theta}_m\}$, hence, it holds for some $\hat{\theta}_{\bar{j}}$ with $d(\hat{\theta}_{\bar{j}}, \theta_0) \leq \varepsilon$. In turn, this implies (by triangle inequality) $d(\hat{\theta}_*, \theta_0) \leq 3\varepsilon$. We conclude that

$$\Pr\left(d(\hat{\theta}_*, \theta_0) > 3\varepsilon\right) \leq \Pr(\mathcal{E}_1).$$

The rest of the proof repeats the argument of part **a** since

$$\Pr(\mathcal{E}_1^c) = \Pr\left(\sum_{j=1}^m I\left\{d(\hat{\theta}_j, \theta_0) > \varepsilon\right\} \geq \frac{m}{2}\right),$$

where \mathcal{E}_1^c is the complement of \mathcal{E}_1 .

APPENDIX B: PROOF OF THEOREM 3.3

By the definition of Wasserstein distance d_{W_1} ,

$$(B.1) \quad d_{W_1, \rho}(\delta_0, \Pi_l(\cdot | \mathcal{X}_l)) = \int_{\Theta} \rho(\theta, \theta_0) d\Pi_l(\theta | X_1, \dots, X_l).$$

(recall that ρ is taken the Hellinger distance). Let R be a large enough constant to be determined later. Note that the Hellinger distance is uniformly bounded by 1. Using (B.1), it is easy to see that

$$(B.2) \quad d_{W_1, \rho}(\delta_0, \Pi_l(\cdot | \mathcal{X}_l)) \leq R\varepsilon_l + \int_{h(P_\theta, P_0) \geq R\varepsilon_l} d\Pi_l(\cdot | \mathcal{X}_l).$$

To this end, it remains to estimate the second term in the sum above. We will follow the proof of Theorem 2.1 in [Ghosal, Ghosh and Van Der Vaart \(2000\)](#). Bayes formula implies that

$$\Pi_l(\theta : h(p_\theta, p_0) \geq R\varepsilon_l | \mathcal{X}_l) = \int_{h(p_\theta, p_0) \geq R\varepsilon_l} \frac{\prod_{i=1}^l \frac{p_\theta(X_i)}{p_0} d\Pi(\theta)}{\int_{\Theta} \prod_{i=1}^l \frac{p_\theta(X_i)}{p_0} d\Pi(\theta)}.$$

Let

$$A_l = \left\{ \theta : -P_0 \left(\log \frac{p_\theta}{p_0} \right) \leq \varepsilon_l^2, P_0 \left(\log \frac{p_\theta}{p_0} \right)^2 \leq \varepsilon_l^2 \right\}.$$

For any $C_1 > 0$, Lemma 8.1 [Ghosal, Ghosh and Van Der Vaart \(2000\)](#) yields

$$\Pr \left\{ \int_{\Theta} \prod_{i=1}^l \frac{p_\theta}{p_0}(X_i) dQ(\theta) \leq \exp \left(- (1 + C_1) l \varepsilon_l^2 \right) \right\} \leq \frac{1}{C_1^2 l \varepsilon_l^2}.$$

for every probability measure Q on the set A_l . Moreover, by the assumption on the prior Π ,

$$\Pi(A_l) \geq \exp \left(- C l \varepsilon_l^2 \right).$$

Consequently, with probability at least $1 - \frac{1}{C_1^2 l \varepsilon_l^2}$,

$$\int_{\Theta} \prod_{i=1}^l \frac{p_\theta}{p_0}(X_i) d\Pi(\theta) \geq \exp \left(- (1 + C_1) l \varepsilon_l^2 \right) \Pi(A_l) \geq \exp \left(- (1 + C_1 + C) l \varepsilon_l^2 \right).$$

Define the event $B_l = \left\{ \int_{\Theta} \prod_{i=1}^l \frac{p_\theta}{p_0}(X_i) d\Pi(\theta) \leq \exp \left(- (1 + C_1 + C) l \varepsilon_l^2 \right) \right\}$.

Let Θ_l be the set satisfying conditions of Theorem 3.3. Then by Theorem 7.1 in Ghosal, Ghosh and Van Der Vaart (2000), there exist test functions $\phi_l := \phi_l(X_1, \dots, X_l)$ and a universal constant K such that

$$(B.3) \quad \begin{aligned} \mathbb{E}_{P_0} \phi_l &\leq 2e^{-Kl\varepsilon_l^2}, \\ \sup_{\theta \in \Theta_l, h(P_\theta, P_0) \geq R\varepsilon_l} \mathbb{E}_{P_\theta} (1 - \phi_l) &\leq e^{-KR^2 \cdot l\varepsilon_l^2}, \end{aligned}$$

where $KR^2 - 1 > K$.

Note that

$$\Pi_l(\theta : h(P_\theta, P_0) \geq R\varepsilon_l | X_1, \dots, X_l) = \Pi_l(\theta : h(p_\theta, p_0) \geq R\varepsilon_l | X_1, \dots, X_l)(\phi_l + 1 - \phi_l).$$

For the first term,

$$(B.4) \quad \mathbb{E}_{P_0} \left[\Pi_l(\theta : h(p_\theta, p_0) \geq R\varepsilon_l | X_1, \dots, X_l) \cdot \phi_l \right] \leq \mathbb{E}_{P_0} \phi_l \leq 2e^{-Kl\varepsilon_l^2}.$$

Next, by the definition of B_l , we have

$$(B.5) \quad \begin{aligned} &\Pi_l(\theta : h(p_\theta, p_0) \geq R\varepsilon_l | \mathcal{X}_l)(1 - \phi_l) = \\ &\frac{\int_{h(p_\theta, p_0) \geq R\varepsilon_l} \prod_{i=1}^l \frac{p_\theta}{p_0}(X_i) d\Pi(\theta) (1 - \phi_l)}{\int_{\Theta} \prod_{i=1}^l \frac{p_\theta}{p_0}(X_i) d\Pi(\theta)} (I\{B_l\} + I\{B_l^c\}) \\ &\leq I\{B_l\} + e^{(1+C_1+C)l\varepsilon_l^2} \int_{h(p_\theta, p_0) \geq R\varepsilon_l} \prod_{i=1}^l \frac{p_\theta}{p_0}(X_i) d\Pi(\theta) (1 - \phi_l). \end{aligned}$$

To estimate the second term of last equation, note that

$$(B.6) \quad \begin{aligned} &\mathbb{E}_{P_0} \int_{h(P_\theta, P_0) \geq R\varepsilon_l} \prod_{i=1}^l \frac{p_\theta}{p_0}(X_i) d\Pi(\theta) (1 - \phi_l) \leq \\ &\mathbb{E}_{P_0} \left(\int_{\theta \in \Theta \setminus \Theta_l} \prod_{i=1}^l \frac{p_\theta}{p_0}(X_i) d\Pi(\theta) (1 - \phi_l) + \int_{\Theta_l \cap h(P_\theta, P_0) \geq R\varepsilon_l} \prod_{i=1}^l \frac{p_\theta}{p_0}(X_i) d\Pi(\theta) (1 - \phi_l) \right) \leq \\ &\Pi(\Theta \setminus \Theta_l) + \int_{\Theta_l \cap h(p_\theta, p_0) \geq R\varepsilon_l} \mathbb{E}_{P_0} \left(\prod_{i=1}^l \frac{p_\theta}{p_0}(X_i) d\Pi(\theta) (1 - \phi_l) \right) \leq \\ &e^{-l\varepsilon_l^2(C+4)} + e^{-KR^2 \cdot l\varepsilon_l^2} \leq 2e^{-l\varepsilon_l^2(C+4)} \end{aligned}$$

for $R \geq \sqrt{(C+4)/K}$, hence the second term in (B.5) is bounded by $2e^{-(3-C_1)l\varepsilon_l^2}$.

Set $C_1 = 1$ and note that $I\{B_l\} = 1$ with probability $P(B_l) \leq 1/l\varepsilon_l^2$. It follows from (B.4), (B.5) and (B.6) and Chebyshev's inequality that for any $t > 0$

$$\begin{aligned} \Pr\left(\Pi_l(\theta : h(p_\theta, p_0) \geq R\varepsilon_l | \mathcal{X}_l) \geq t\right) &\leq \Pr(B_l) + \frac{2e^{-Kl\varepsilon_l^2}}{t} + \frac{2\exp(-2l\varepsilon_l^2)}{t} \\ &\leq \frac{1}{l\varepsilon_l^2} + \frac{2e^{-Kl\varepsilon_l^2}}{t} + \frac{2\exp(-2l\varepsilon_l^2)}{t}. \end{aligned}$$

Finally, (B.3) implies that, for a constant $\tilde{K} = \min(K/2, 1)$ and $t = e^{-\tilde{K}l\varepsilon_l^2}$,

$$\begin{aligned} \Pr\left(\Pi_l(\theta : h(p_\theta, p_0) \geq R\varepsilon_l | \mathcal{X}_l) \geq t\right) &\leq \frac{1}{l\varepsilon_l^2} + 2e^{-Kl\varepsilon_l^2/2} + 2\exp(-l\varepsilon_l^2) \\ &\leq \frac{1}{l\varepsilon_l^2} + 4e^{-(1+K/2)l\varepsilon_l^2}. \end{aligned}$$

Recall that conditions of Theorem 3.3 imply that $t = e^{-\tilde{K}l\varepsilon_l^2} \leq \varepsilon_l$, hence (B.2) gives the final result.

APPENDIX C: PROOF OF THEOREM 3.14

The proof strategy is similar to Theorem 3.3. Note that

$$(C.1) \quad d_{W_{1,\rho}}(\delta_0, \Pi_{n,m}(\cdot | \mathcal{X}_l)) \leq R\varepsilon_l + \int_{h(P_\theta, P_0) \geq R\varepsilon_l} d\Pi_l(\cdot | \mathcal{X}_l).$$

Let $\mathcal{E}_l := \{\theta : h(P_\theta, P_0) \geq R\varepsilon_l\}$. By the definition of $\Pi_{n,m}$, we have

$$(C.2) \quad \Pi_{n,m}(\mathcal{E}_l | \mathcal{X}_l) = \frac{\int_{\mathcal{E}_l} \left(\prod_{j=1}^l \frac{p_\theta}{p_0}(X_j)\right)^m d\Pi(\theta)}{\int_{\Theta} \left(\prod_{j=1}^l \frac{p_\theta}{p_0}(X_j)\right)^m d\Pi(\theta)}.$$

To bound the denominator from below, we proceed as before. Let

$$\Theta_l = \left\{ \theta : -P_0 \left(\log \frac{p_\theta}{p_0} \right) \leq \varepsilon_l^2, P_0 \left(\log \frac{p_\theta}{p_0} \right)^2 \leq \varepsilon_l^2 \right\}.$$

Let B_l be the event defined by

$$B_l := \left\{ \int_{\Theta_l} \left(\prod_{i=1}^l \frac{p_\theta}{p_0}(X_i) \right)^m dQ(\theta) \leq \exp(-2ml\varepsilon_l^2) \right\},$$

where Q is a probability measure supported on Θ_l . Lemma 8.1 in [Ghosal, Ghosh and Van Der Vaart \(2000\)](#) yields that $\Pr(B_l) \leq \frac{1}{l\varepsilon_l^2}$ for any Q , in particular, for the conditional distribution $\Pi(\cdot|\Theta_l)$. We conclude that

$$\int_{\Theta} \left(\prod_{j=1}^l \frac{p_\theta}{p_0}(X_j) \right)^m d\Pi(\theta) \geq \Pi(\Theta_l) \exp(-2ml\varepsilon_l^2) \geq \exp(-(2m + C)l\varepsilon_l^2).$$

To estimate the numerator in (C.2), note that if Theorem 3.13 holds for $\gamma = \varepsilon_l$, then it also holds for $\gamma = L\varepsilon_l$ for any $L \geq 1$. This observation implies that

$$\sup_{\theta \in \mathcal{E}_l} \left(\prod_{j=1}^l \frac{p_\theta}{p_0}(X_j) \right)^m \leq e^{-c_1 R^2 m l \varepsilon_l^2}$$

with probability $\geq 1 - 4e^{-c_2 R^2 l \varepsilon_l^2}$, hence

$$\int_{\mathcal{E}_l} \left(\prod_{j=1}^l \frac{p_\theta}{p_0}(X_j) \right)^m d\Pi(\theta) \leq e^{-c_1 R^2 m l \varepsilon_l^2}$$

with the same probability. Choose $R = R(C)$ large enough so that $c_1 m R^2 \geq 3m + C$. Putting the bounds for the numerator and denominator of (C.2) together, we get that with probability $\geq 1 - \frac{1}{l\varepsilon_l^2} - 4e^{-c_2 R^2 l \varepsilon_l^2}$,

$$\Pi_{n,m}(\mathcal{E}_l|\mathcal{X}_l) \leq e^{-ml\varepsilon_l^2} \leq \varepsilon_l,$$

where the last inequality follows from our assumptions on ε_l . The result now follows from (C.1).

APPENDIX D: PROBABILISTIC PARAFAC MODEL

The generative model using p-parafac is defined as follows. First, p-parafac generates a discrete random measure $\nu(\cdot) = \sum_{h=1}^{\infty} \nu_h \delta_h(\cdot)$ using the stick-breaking construction of DP ([Sethuraman, 1994](#)) as

$$V_h \sim \text{Beta}(1, \alpha) \text{ and } \nu_h = V_h \prod_{l < h} (1 - V_l) \text{ for } h = 1, \dots, \infty,$$

where ν_h represents the prior probability of responders responses belonging to the latent class h . Given ν , p-parafac then samples the latent class z_n from ν for responder $n = 1, \dots, 4067$. For each latent class h and categories $s \in \{\text{Abort}, \text{Mar}, \text{Cap}\}$, p-parafac also generates a sample in the two-dimensional unit simplex $\Psi_h^s = (\psi_{h;\text{yes}}^s, \psi_{h;\text{no}}^s)$ from $\text{Dirichlet}(\alpha_1, \alpha_2)$. Finally, given Ψ_h^g 's and z_n 's, the response of n -th responder in category s , $y_{n,s}$, is modeled as

$$y_{n,s} \sim \text{Multinomial}(\{\text{yes}, \text{no}\}, \Psi_{z_n}^s).$$

This generative model in turn implies that

$$\pi_{amc} = \sum_{h=1}^{\infty} \nu_h \Psi_{h;amc}, \text{ where } \Psi_{h;amc} = (\psi_{h;a}^{\text{Abort}}, \psi_{h;r}^{\text{Mar}}, \psi_{h;c}^{\text{Cap}}),$$

$a \in \{\text{yes}, \text{no}\}$, $r \in \{\text{yes}, \text{no}\}$, $c \in \{\text{yes}, \text{no}\}$, and ν_h 's and ψ_h^s 's respectively have the stick-breaking and Dirichlet prior distributions. The hyperparameters of this model are α , α_1 , and α_2 . We specify Gamma prior on α with scale and shape parameter fixed at 1 and assume that $\alpha_1 = 1$ and $\alpha_2 = 1$. Due to the finite number of available responses, the number of latent classes is upper-bounded by a finite number ([Ishwaran and James, 2001](#)). This formulation leads to a simple Gibbs sampler for obtaining posterior samples of π_{amc} (see (5) in [Dunson and Xing \(2009\)](#) for analytic forms of the conditional distributions).

STANISLAV MINSKER
MATHEMATICS DEPARTMENT
DUKE UNIVERSITY, Box 90320
DURHAM, NC 27708-0320
E-MAIL: sminsker@math.duke.edu

SANVESH SRIVASTAVA, LIZHEN LIN AND DAVID B. DUNSON
DEPARTMENT OF STATISTICAL SCIENCE
DUKE UNIVERSITY, Box 90251
DURHAM NC 27708-0251
E-MAIL: ss602@stat.duke.edu
lizhen@stat.duke.edu
dunson@duke.edu

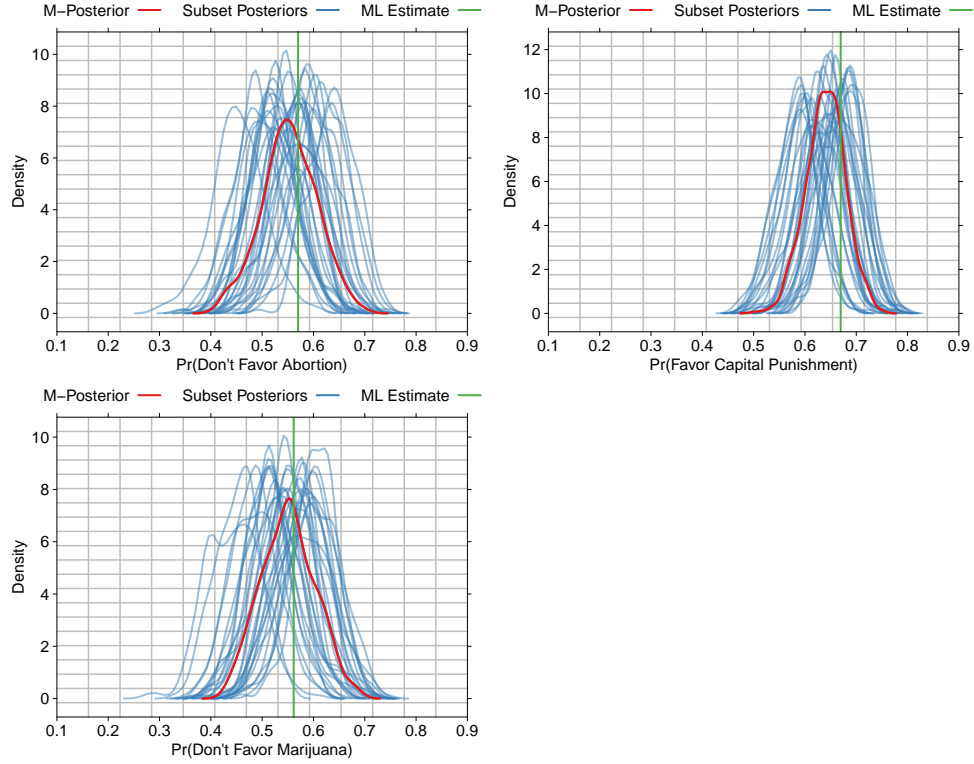


Fig 4: Subset posteriors and M-posteriors of Abort, Mar, and Cap. The x axis represents the posterior draws of marginal probabilities of Abort ($\pi_{\text{no}\bullet\bullet}$), Mar ($\pi_{\bullet\text{no}\bullet}$), and Cap ($\pi_{\bullet\bullet\text{yes}}$), and the curves represent the corresponding kernel density estimators. Density of M-posterior is the red curve. Note that the maximum likelihood estimator of these probabilities (calculated using the whole sample) are very close to the mode of the M-posterior.

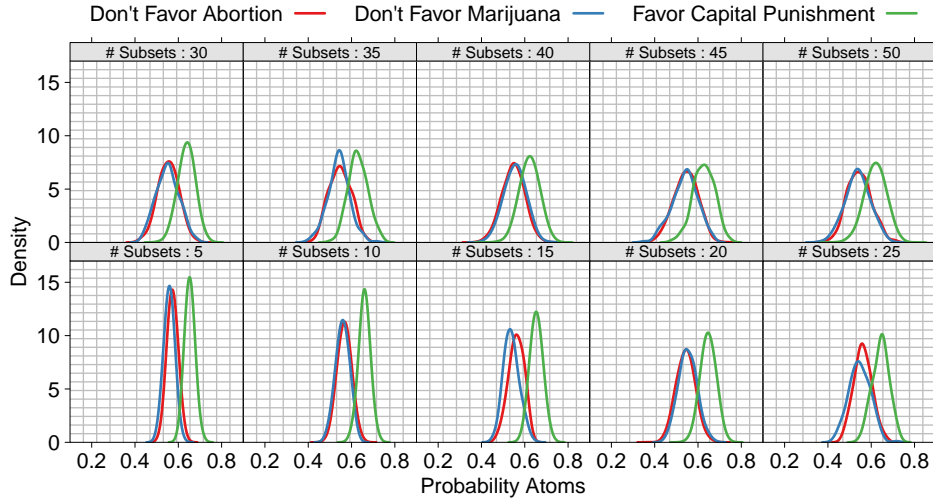


Fig 5: Effect of the number of subsets on M-posterior. The number of subsets (m) varies along the panels. In each panel, the x axis represents the posterior draws of marginal probabilities of Abort ($\pi_{\text{no}\bullet\bullet}$), Mar ($\pi_{\bullet\text{no}\bullet}$), and Cap ($\pi_{\bullet\bullet\text{yes}}$). The curves represent the corresponding kernel density estimators. The rate of concentration of M-posterior around its mode slowly decreases as m increases. The location of the mode of M-posterior, however, is stable and does not fluctuate with increasing m . The heuristic approach explained in remark 4.1 (see equation (4.1)) suggests the “optimal” values $m_* = 35$ (Abort), $m_* = 35$ (Cap), $m_* = 25$ (Mar).